Intron Retention Controls Localization of lncRNAs *PURPL* and *MALAT1* to Promote Cell Proliferation and Migration

Ioannis Grammatikakis[1], Chosita Norkaew[1], You Jin Song[2], Amit K. Behera[3], Erica C. Pehrsson[4,5], Corrine Corrina R. Hartford[1], Shreya Kordale[3], Rishabh Prasanth[2], Yongmei Zhao[6], Biraj Shrethsa[6], Xiao Ling Li[1], Ravi Kumar[1], Ragini Singh[1], Tayvia Brownmiller[7], Xinyu Wen[4,5], Natasha J. Caplen[7], Pablo Perez-Pinera[8], Kannanganattu V. Prasanth[2*], Thomas Gonatopoulos-Pournatzis[3*], Ashish Lal[1*]

[1]Regulatory RNAs and Cancer Section, Genetics Branch, Center for Cancer Research (CCR), National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, Maryland, USA

[2]Department of Cell and Developmental Biology, Cancer Center at Illinois, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[3]Functional Transcriptomics Section, RNA Biology Laboratory, NCI, NIH, Frederick, Maryland, USA

[4]OMICS Technology Facility, Bioinformatics, Genetics Branch, CCR, NCI, NIH, Bethesda, Maryland, USA

[5]Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA

[6]Sequencing Facility Bioinformatics Group, Bioinformatics and Computational Science Directorate, Frederick National Laboratory for Cancer Research, NCI, NIH, Frederick, Maryland, USA

[7]Functional Genetics Section, Genetics Branch, CCR, NCI, NIH, Bethesda, Maryland, USA

[8]Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

*Corresponding authors: ashish.lal@nih.gov; kumarp@illinois.edu; thomas.gonatopoulos@nih.gov

Classification: Biological Sciences

Keywords: U2AF2, lncRNA, *PURPL*, *MALAT1*, intron retention, splicing, CRISPR/Cas9, nuclear speckles, proliferation, cell migration

Abstract

Intron retention (IR) is a form of alternative splicing in which introns that are normally removed are retained in mature transcripts. Despite emerging evidence of widespread IR in protein-coding genes and lncRNAs, the mechanisms and functional consequences underlying this process remain poorly understood. Here, we performed a genome-wide screen, to dissect the mechanisms governing IR in the lncRNA *PURPL.* Unexpectedly, the top hit from the screen was the essential splicing activator U2AF2, which promotes IR in *PURPL* through direct binding to a weak polypyrimidine tract. Retention of this intron drives nuclear localization of *PURPL* and enhances cell proliferation, revealing a functional role for IR. Transcriptome-wide analysis showed that while U2AF2 promotes splicing of most transcripts, consistent with its canonical role, it also promotes IR in a distinct subset of RNAs. This subset includes the nuclear speckle localized lncRNA *MALAT1*, whose speckle localization is impaired upon U2AF2 depletion. Using *MALAT1* knockout cells reconstituted with wild-type or intron-deleted *MALAT1* variants, we identified a single intron that is essential for *MALAT1* nuclear speckle localization. Deletion of this intron from endogenous *MALAT1* disrupted speckle localization and reduced cell migration, phenocopying the loss of *MALAT1*. Together, these findings uncover a previously unrecognized role for U2AF2 in promoting intron retention and establish IR as a key mechanism regulating lncRNA localization and function.

70 **Significance statement**

71 Intron retention is an understudied form of alternative splicing that can profoundly influence
72 protein isoform diversity, RNA stability, localization, and function. We investigated the
73 mechanisms and consequences of intron retention in the lncRNAs *PURPL* and *MALAT1*. Using a
74 genome wide CRISPR screen and transcriptome wide analyses, we discovered that the splicing
75 factor U2AF2, classically known for promoting intron removal, unexpectedly enhances intron
76 retention by binding to a weak polypyrimidine tract in the retained intron of these lncRNAs. Intron
77 retention, in turn, dictates their localization and modulates cell proliferation and migration,
78 revealing a previously unrecognized regulatory layer in lncRNA biology.

79

80

81

82

83

84

85

86

87

88

89

90

91

92

Introduction

Splicing refers to the process of intron removal and exon "stitching" as part of RNA processing that occurs in the nucleus and is catalyzed by a large nucleoprotein complex called the spliceosome (1-3). Splicing is a highly regulated process where *cis*- and *trans*-acting factors coordinate proper and efficient spliceosomal recruitment (4). Several consensus sequences in the RNA are crucial for splicing: the 5′ and 3′ splice sites (ss), which demarcate the exon-intron boundaries of the intron to be removed, typically conform to the consensus sequence of GU-AG. The polypyrimidine (Py)-tract is a stretch of pyrimidines close to the 3′ ss and the branch point that lies upstream of the Py-tract (1). The splicing factor U2AF1 (U2-auxiliary factor 1) binds to the 3′ ss and U2AF2 binds to the Py-tract. The U2AF heterodimer, along with SF1 (Splicing factor 1), is recruited to the branch point to form the pre-spliceosomal E complex. This complex serves as a critical intermediate for the subsequent assembly of the spliceosome and splicing catalysis (5-9).

The vast majority of human genes encode pre-mRNAs that contain introns that are spliced out and most nascent transcripts undergo alternative splicing, which leads to the combinatorial joining of exons. Alternative splicing is a major contributor to protein isoform diversity, enabling the generation of multiple transcript variants from a single gene and frequently shaping measurable cellular phenotypes (10-13). There are several types of alternative splicing such as exon skipping, intron retention, mutually exclusive exons, and alternative 5′ and 3′ ss (14). One of the least studied forms of alternative splicing is intron retention (IR). Lately, it has drawn more attention since it has an important role in post-transcriptional gene expression regulation, and it is believed that more than half of human multi-exonic genes are affected by IR (15). IR frequently leads to nuclear detention of transcripts and eventually degradation which may serve as a defense mechanism of the cell against deleterious products. The extra sequence added by the retained intron may change

116 the open reading frame (ORF) of protein-coding genes and lead to nonsense-mediated decay
117 (NMD) (16-18) or initiate the use of an upstream ORF (19, 20). In addition, they can contain
118 binding sites for microRNAs (miRNAs) or RNA-binding proteins (RBPs), or they may form
119 structural changes that impact the function of the transcript (21). Interestingly, IR has been
120 associated with several diseases such as cancer where tumor cells show higher levels of IR
121 compared to normal tissues (22). Studies on IR have shown that retained introns have been
122 associated with weaker (non-consensus) splice sites, higher GC content, various lengths of introns,
123 and higher intronic sequence conservation (15, 23). However, the molecular mechanisms of IR
124 regulation remain largely elusive.

125 Long non-coding RNAs (lncRNAs) are RNA molecules >200 nucleotides (nt) long that
126 have no obvious coding potential. Similar to protein-coding genes, lncRNAs are transcribed by
127 RNA Polymerase II, are 5′-capped, spliced and polyadenylated (24). However, several
128 characteristics distinguish lncRNAs from protein-coding genes, including higher tissue-specific
129 expression, lower abundance and stability, predominant nuclear localization, and having fewer but
130 longer exons (25-29). One of the most distinguished features of lncRNAs is that they are less
131 efficiently spliced compared to mRNAs (27, 30-32). Interestingly, lncRNAs contain introns with
132 much slower intron-excision kinetics compared to protein-coding genes (33). However, the
133 regulatory mechanisms governing lncRNA splicing remain poorly understood.

134 lncRNAs have been shown to participate in splicing regulation, with *MALAT1* as the most
135 prominent example. *MALAT1* is a highly abundant nuclear lncRNA which has oncogenic activity
136 and is one of the first lncRNAs associated with cancer (34-36). It also has the ability to regulate
137 alternative splicing because it resides in the nuclear speckles and can interact with several SR
138 (Serine/Arginine rich) proteins to regulate their phosphorylation state and distribution between

139 transcription sites and speckles (35, 37-39). Nuclear speckles are sub-nuclear compartments that
140 act as hubs in coordinating transcription, RNA processing and mRNA export (40). Several RBPs
141 reside in nuclear speckles such as the protein SON which is a key component and marker of nuclear
142 speckles (41). However, *MALAT1* itself is not essential for the structural integrity of nuclear
143 speckles (42, 43). Nevertheless, it remains unclear whether *MALAT1* undergoes alternative
144 splicing and how this process impacts its localization to nuclear speckles.

145 Another lncRNA that we previously characterized and termed *PURPL* is transcriptionally
146 induced by p53 (44). *PURPL* plays a growth-promoting role in colorectal and liver cancer cells by
147 distinct mechanisms including suppressing p53 protein levels. Interestingly, transcriptomic
148 analyses using both long-read and short-read sequencing revealed multiple *PURPL* isoforms,
149 several of which retain intron 2 (45). However, the molecular mechanism governing *PURPL* intron
150 2 splicing has yet to be elucidated.

151 Here, we show that the intron 2-retaining *PURPL* isoform, which localizes to the nucleus,
152 has a shorter half-life than the spliced isoform and promotes cell proliferation. To identify
153 regulatory factors controlling the splicing of this functionally relevant intron, we utilized our
154 recently developed CRISPR-based genome-wide screening strategy using a pooled guide RNA
155 library with a *PURPL* intron 2 minigene reporter containing its flanking exons. Unexpectedly, this
156 approach revealed a role of the splicing activator U2AF2in promoting *PURPL* IR, contrary to its
157 canonical function in facilitating intron removal. U2AF2 was further found to promote a broader
158 intron retention program that includes two introns within *MALAT1*. Notably, depletion of U2AF2
159 resulted in *MALAT1* exclusion from nuclear speckles. *MALAT1* knockout cells exogenously
160 expressing *MALAT1* truncation mutants lacking specific introns identified one intron essential for
161 the nuclear speckle association of the *MALAT1* transcript. Breast cancer cells where this intron is

162 deleted showed reduced cell migration demonstrating a functional role of the intronic sequences.
163 These findings provide mechanistic insights into the splicing regulation of *PURPL* and *MALAT1*,
164 uncovering a previously unrecognized non-canonical function of U2AF2 in promoting
165 functionally relevant intron retention events.

166 **Results**

167 **A CRISPR-based screen identifies U2AF2 as an activator of *PURPL* intron 2 retention**

168 We previously identified *PURPL* as a p53-regulated lncRNA that promotes the growth of
169 colorectal and liver cancer cells (44, 45). In those studies, we reported that *PURPL* undergoes
170 alternative splicing, with the ~1.9 kb intron 2 either retained or spliced in the mature transcript
171 (45). To identify regulators of *PURPL* intron 2 retention, we applied our recently developed
172 screening platform, CRISPR-based identification of Regulators of Alternative Splicing with
173 Phenotypic Sequencing (CRASP-seq) (46). This approach utilizes a genome-wide knockout
174 lentiviral library targeting ~18,500 protein-coding genes using a dual hybrid guide (hg)RNA
175 system that combines *Streptococcus pyogenes* (*Sp*)Cas9 and optimized *Acidaminococcus sp.*
176 (op)Cas12a nucleases. (13, 47-49). The library also includes ~2,400 pairs and ~500 curated genetic
177 interaction pairs, allowing combinatorial perturbations from a single construct. Each gene or gene
178 pair is targeted by four independent hgRNAs alongside ~500 intergenic and non-targeting controls,
179 totaling ~ 91,000 hgRNAs. The lentiviral library incorporates a doxycycline-inducible *PURPL* IR
180 reporter comprised of the full native intron 2 sequences (~1.9 kb) flanked by the adjacent exons 2
181 (161 bp) and 3 (163 bp) (Figure 1A).

182 We performed the CRASP-screen in two cell lines: HAP1 and RPE1. Following puromycin
183 selection for 48 h and doxycycline induction for an additional 24 h, RNA was extracted, and poly-
184 A RNA was purified using oligo(dT) beads to exclude transcripts still undergoing processing.

185 cDNA was synthesized using customized primers containing unique molecular identifiers (UMIs),
186 facilitating the removal of PCR duplicates during downstream analysis. The synthesized cDNA
187 was then utilized to construct Illumina sequencing libraries for paired-end sequencing. This
188 sequencing approach enables simultaneous quantification of transcripts that either retain or splice
189 out an intron of interest from one read, while the corresponding hgRNA sequences were captured
190 in the other read to pinpoint specific genetic perturbations. By directly linking *PURPL* splicing
191 outcomes with genetic modifications introduced by each hgRNA in the library, this platform
192 allowed the genome-wide identification of *PURPL* splicing regulators (Figure 1B).

193 We employed an analytical pipeline that prioritized candidate genes based on multiple
194 hgRNAs showing significant deviations from intergenic controls, with consistent identification
195 across two technical replicates. This analysis identified 62 genes that influence *PURPL* intron 2
196 splicing, the majority of which act as suppressors of intron 2 splicing (Figure 1C and Table S1).
197 Gene ontology analysis of biological processes revealed a strong enrichment for terms related to
198 mRNA processing and RNA splicing. Additionally, analysis of cellular components indicated that
199 the identified hits are predominantly associated with Spliceosomal Complex, including the U2-
200 type Spliceosomal Complex and the Catalytic step 2 Spliceosome (Figure 1D). These findings
201 validate the effectiveness of our screening approach.

202 Consistent with the endogenous *PURPL* transcript isoforms, the basal levels of IR of the
203 reporter were very high, with the majority of amplified transcripts corresponding to intron-
204 containing isoforms, as expected based on the intensity of the Illumina libraries before sequencing
205 (Figure S1A). Thus, the detection of genes that suppress IR might be below the detection levels of
206 this screen. Therefore, we focused on the genes that promote intron retention. By calculating the
207 change in Percent Intron Retention ($\Delta PIR = PIR_{guide} - PIR_{intergenic}$) across the two cell lines, we

208 identified 5 unique genes as regulators of *PURPL* intron 2 retention, all of which act to suppress
209 intron 2 splicing (Figure 1E and Table S1). This highly significant overlap between the two cell
210 lines (Figure 1E; odds ratio = 204.4, $p=2.81xe^{-10}$; Fisher's exact test) indicated a shared regulatory
211 mechanism governing *PURPL* IR, although cell-type-specific effects may also contribute.

212 Surprisingly, the gene with the strongest impact based on the average ΔPIR metric and
213 identified as a hit in both cell lines, was U2AF2 (Figure 1F). This was followed by Small Nuclear
214 Ribonucleoprotein U1 Subunit 70 (SNRNP70), a core component of the U1 snRNP, and
215 Elongation Factor Tu GTP Binding Domain, which is associated with U5 snRNP (Figure 1F). Our
216 finding that U2AF2 promotes IR of *PURPL* intron 2 was unexpected because U2AF2 is a
217 component of the spliceosome machinery, and its canonical function is to promote intron removal
218 (5, 9).

219 **U2AF2 binds to the *PURPL* intron 2 transcript and promotes intron retention**

220 U2AF2, as a heterodimer with U2AF1, binds to the Py-tract and the 3′ ss, respectively, to promote
221 splicing (5, 8). To elucidate its role in *PURPL* splicing, we investigated whether U2AF2 binds to
222 intron 2. Enhanced Crosslinking and Immunoprecipitation sequencing (eCLIP-seq) (50), from the
223 ENCODE project (51, 52) from HepG2 cells revealed multiple U2AF2 and U2AF1 binding sites
224 not only proximal to the 3′ ss of *PURPL* intron 2 but also broadly distributed across intron 2 (Figure
225 2A). Compared to eCLIP data for other RBPs, U2AF1 and U2AF2 showed the highest density of
226 binding sites and the most prominent peaks. Specific binding of U2AF2 to intron 2 was confirmed
227 by performing RNA immunoprecipitation (RNA-IP) using a U2AF2-specific antibody and primer
228 pairs designed to detect *PURPL* transcripts. U2AF2 was found to interact with transcripts
229 containing *PURPL* intron 2, but not with the spliced *PURPL* transcripts (Figure 2B and S1B).

230 To functionally validate the role of U2AF2 in *PURPL* intron 2 retention, we performed
231 RT-PCR assays from HAP1 and RPE1 cells transduced with *PURPL* minigene reporter constructs
232 and control or *U2AF2*-targeting sgRNAs. In both cell lines, we observed a striking decrease in
233 intron 2 retention upon *U2AF2* depletion as compared to control cells expressing sgRNAs targeting
234 intergenic regions (Figure 2C). As an alternative approach, we knocked down *U2AF2* using two
235 independent siRNAs and conducted semi-quantitative RT-PCR assays from SKHEP1, HepG2, and
236 normal human diploid fibroblasts W138 cells, using a primer triplet (Figures 2D-2F; Figures S2A
237 and S2B). In all tested cell lines, U2AF2 depletion increased *PURPL* intron 2 splicing, from ~20%
238 to ~40-60%, more than doubling the abundance of the spliced isoform. Given that U2AF2
239 functions as a heterodimer with U2AF1, we also knocked down *U2AF1* in SKHEP1 cells and
240 observed an effect similar to *U2AF2* knockdown on *PURPL* intron 2 retention regulation (Figure
241 S2C). We expected that U2AF1 would also come as one of the top hits in the screen but *U2AF1*
242 sgRNAs were not included in the library due to high off-target scores of the guides. These results
243 demonstrate that the U2AF complex promotes *PURPL* intron 2 retention and that both U2AF1 and
244 U2AF2 promote intron retention in *PURPL*.

245 **The intron-retained *PURPL* transcript is unstable and localizes to the nucleus**

246 We previously observed an enrichment of the intron 2-retaining *PURPL* transcript in the nucleus
247 in liver cancer cells (45). To validate these findings, we conducted single molecule RNA-FISH in
248 HCT116 cells using probes hybridizing to *PURPL* intron 2 (Figure 3A). Because *PURPL* is
249 upregulated in response to DNA damage (44), we treated the cells with hydroxyurea for 24 h to
250 induce *PURPL* expression. RNA-FISH revealed that the *PURPL* intron 2 containing transcript was
251 exclusively localized to the nucleus. To assay for subnuclear compartmentalization, we also
252 performed RNA-FISH to detect *MALAT1*, which resides in nuclear speckles (38). Intron 2-retained

253 *PURPL* transcripts did not colocalize with *MALAT1* indicating that *PURPL* is excluded from
254 nuclear speckles (Figure 3A). We further validated this result by conducting nuclear/cytoplasmic
255 fractionation assays followed by RT-qPCR and found that *PURPL* intron 2-containing transcripts
256 were almost exclusively nuclear (Figure S3A).

257 Intron retention can affect transcript stability, with longer RNA species often being more
258 vulnerable to degradation. To test if *PURPL* intron 2 effects transcript turnover, we conducted
259 RNA stability assays using the transcription inhibitor Actinomycin D (ActD) and observed that
260 the intron 2-retained *PURPL* transcript was less stable and had a significantly lower half-life than
261 the *PURPL* transcript in which intron 2 was spliced out (Figure 3B). To determine the functional
262 impact of *PURPL* intron 2 retention, we used CRISPRi to generate *PURPL*-depleted SKHEP1
263 cells by targeting 3 independent sgRNAs to the p53-response element of the *PURPL* promoter that
264 we had used in our previous study (45). We observed >80% reduction in the levels of endogenous
265 *PURPL* in these experiments (Figure 3C). We next stably transduced these *PURPL*-depleted cells
266 with a lentivirus to express exogenous *PURPL* transcript specifically containing intron 2 under the
267 control of a doxycycline (doxy)-inducible promoter and observed ~2-3-fold induction of the
268 exogenous *PURPL* intron 2 containing transcript (Figure 3D). Because we and others have
269 previously shown that *PURPL* is overexpressed in some cancers and *PURPL* promotes
270 proliferation (44, 45, 53, 54), we next performed cell proliferation assays from the doxy-inducible
271 cells. Our results showed that cells expressing intron 2-contaning *PURPL* proliferate faster (~2-
272 fold) as compared to the no-doxy cells after 6 days of doxy treatment (Figure 2E). Induction of
273 intron 2-contaning *PURPL* transcripts did not have any effect on live cell percentage which
274 remained at ~100% (Figure S3B). Importantly, as a negative control, there was no effect of doxy
275 treatment on the proliferation of parental SKHEP1 cells (Figure S3C). To confirm that

276 proliferation increase induced by exogenous *PURPL* transcript containing intron 2 was not due to
277 the spliced isoform, we performed semi-quantitative RT-PCR. As shown in Figure 3F, doxy
278 treatment resulted exclusively in the detection of the intron-retained *PURPL* transcript. These data
279 suggest that intron 2 retention modulates *PURPL* function by sequestering the transcript in the
280 nucleus and reducing its stability, while simultaneously promoting cell proliferation, highlighting
281 a critical regulatory role for intron 2.

282 **A weak polypyrimidine tract is critical for *PURPL* intron 2 retention**

283 U2AF2 binds to the Py-tract to initiate spliceosomal assembly leading to splicing and intron
284 removal (5). Suboptimal splicing efficiency may be the result of multiple factors, including splice
285 site strength and the Py-tract sequence composition. The Py-tract is a stretch of 15-20 pyrimidines
286 (mostly Us) immediately upstream of the 3′ ss. We analyzed the Py-tract of *PURPL* intron 2 and
287 noticed that it deviates from the consensus, containing numerous purines (Figure S4A). To
288 investigate whether U2AF2 regulation on intron 2 retention relies on the Py-tract sequence, we
289 used the *PURPL* intron 2 minigene reporter system (Figure 1A). We mutated the Py-tract,
290 replacing it either with a strong Py-tract from *PURPL* intron 1 or with a stretch of Us, simulating
291 a strong Py-tract. *U2AF2* was knocked down in SKHEP1 cells stably expressing the mutated Py-
292 tract constructs and RT-PCR was conducted to monitor intron 2 splicing. Interestingly, the basal
293 levels of intron retention were very different with different mutants; ~90% in intron 2 WT (wild-
294 type), ~30% in *PURPL* intron 1 and ~20% in U-rich construct (Figure S4B and S4C). This data
295 demonstrates that the weak Py-tract is responsible for the inefficient intron 2 removal in *PURPL*
296 (Figure S4C). *U2AF2* knockdown resulted in modest increase in splicing of the mutant constructs
297 as well. Because the minigene reporter contained the mutant Py-tract along with the remaining
298 endogenous sequences from *PURPL* intron 2, the modest increase in splicing observed for the

299 mutants after *U2AF2* knockdown may reflect U2AF2 inhibitory binding at additional sites within
300 *PURPL* intron 2 beyond the Py-tract (Figure 2A). These data indicate that the endogenous intron
301 2 Py-tract in *PURPL* is a non-canonical, weak Py-tract that controls the high intron retention
302 observed in *PURPL*.

303 **U2AF2 directly regulates multiple IR events including *MALAT1***

304 To determine if U2AF2 promotes IR in additional genes, we next performed total RNA-seq and
305 PacBio Iso-seq from SKHEP1 cells transfected with siCTRL or siU2AF2. Consistent with our
306 data, *U2AF2* knockdown resulted in a marked increase in the splicing efficiency of *PURPL* intron
307 2 (Figures S5A and S5B). Additionally, *U2AF2* depletion increased overall *PURPL* expression,
308 with increased read coverage across all exons (Figures S5A and S5B; Tables S3 and S4). To
309 identify IR events regulated by U2AF2 at a transcriptome-wide level, we utilized IRFinder, an
310 algorithm specifically designed to calculate changes in IR (55). This analysis revealed ~1000
311 statistically significant IR events regulated by U2AF2. For majority of these events (732), there
312 was an increase in the IR ratio (Intronic abundance/(Intronic abundance + exonic abundance)) after
313 *U2AF2* depletion, indicating that U2AF2 predominantly promotes splicing, as expected. However,
314 in 269 events, *U2AF2* knockdown resulted in decreased IR, indicating that U2AF2 promotes IR in
315 a subset of genes (Figure 4A and Table S2). We next analyzed ENCODE U2AF2 eCLIP-seq data
316 from HepG2 cells to determine if U2AF2 directly binds to the regulated transcripts. U2AF2
317 binding was detected in 96 out of the 269 IR events supporting a direct role for U2AF2 in
318 promoting IR (Figure S5C and Table S2).

319 To validate these findings, we selected three of the identified introns from *MDM1*,
320 *RETREG2*, and *TMEM41B* (Figure S6A). We observed multiple U2AF2 binding peaks within the
321 regulated introns from *MDM1* (Figure S6B). We next assessed their splicing in SKHEP1 cells

322 following *U2AF2* knockdown using two independent siRNAs. In all three cases, *U2AF2* depletion
323 significantly enhanced intron removal (Figures S6C-S6E). IRFinder validated that *PURPL* intron
324 2 was one of the events showing a significant decrease (~2-fold) in IR ratio upon *U2AF2* depletion
325 (Figure 4B).

326 Among the lncRNAs in which IR is promoted by U2AF2, was the lncRNA *MALAT1*.
327 IRFinder identified two introns in *MALAT1* that have weak Py-tract sequences similar to *PURPL*
328 intron 2 (Figure S7A-S7C). For both introns, *U2AF2* knockdown resulted in a significant decrease
329 (~2-fold) in the IR ratio compared to siCTRL (Figures 4A and 4B). This result was confirmed by
330 RT-PCR from HCT116 and SKHEP1 cells upon *U2AF2* knockdown (Figures 4C-4F).

331 **Intron 2 of *MALAT1* drives its localization to nuclear speckles**

332 *MALAT1* regulates splicing by interacting with splicing factors in nuclear speckles.(38) To
333 determine the impact of IR on *MALAT1* localization to nuclear speckles, we performed RNA-FISH
334 for *MALAT1* upon *U2AF2* depletion in SKHEP1 and HCT116 cells. Remarkably, *U2AF2*
335 knockdown resulted in a significant decrease in the ratio of *MALAT1* signal intensity in speckles
336 relative to nucleoplasm (Figures 5A and 5B; Figures S8A and S8B). U2AF2 depletion did not
337 affect total *MALAT1* levels as observed in our RNA-seq (Table S3). This reduction in *MALAT1*
338 levels in nuclear speckles did not compromise the structural integrity of nuclear speckles as
339 indicated by immunostaining for the RNA-binding protein SON, a well-established nuclear
340 speckle marker (41) (Figures 5A and S8A) and is consistent with previous reports showing that
341 *MALAT1* is not essential for the integrity of nuclear speckles (42, 43). These data suggest that
342 U2AF2 plays an important role in promoting *MALAT1* localization to nuclear speckles.

343 To determine whether it is the intron retention in the transcript *per se* which directs
344 *MALAT1* localization or if U2AF2 indirectly regulates its localization, we assessed the nuclear

345 speckle localization of wild-type (WT) and intron-deleted *MALAT1* transcripts. To do this, we
346 used CRISPR/Cas9 to generate two *MALAT1* knockout (KO) clones in HCT116 by inserting ~100
347 bp downstream of the transcription start site, a puromycin resistance gene in one allele and a
348 hygromycin resistance gene in the other allele. In both KO clones, >95% reduction in *MALAT1*
349 RNA levels was observed as compared to a wild-type clone (Figure S9A).

350 Next, we transduced the *MALAT1* KO clones with an empty vector or doxycycline-
351 inducible lentiviral constructs that exogenously express full-length *MALAT1*-WT in which both
352 introns are intact, intron 1 deleted *MALAT1* (del-1), intron 2 deleted *MALAT1* (del-2) or both
353 introns were deleted (del-1+2). After 48 h of doxy-treatment to induce *MALAT1* expression, which
354 was confirmed with RT-qPCR (Figure S9B), RNA-FISH combined with immunofluorescence for
355 the nuclear speckle marker SON, revealed distinct localization patterns. As expected, exogenously
356 expressed *MALAT1*-WT predominantly (~80%) localized to nuclear speckles (Figures 6A and
357 S10A). Similarly, *MALAT1*-del-1 also predominantly (~80%) localized to nuclear speckles. In
358 contrast, *MALAT1* lacking intron 2 (del-2) or both introns (del-1+2) exhibited a marked reduction
359 (<15%) in nuclear speckle localization and instead displayed diffused localization in the nucleus
360 (Figures 6A and S10A). Importantly, this diffused localization upon intron 2 deletion was observed
361 in both *MALAT1* knockout clones and was further confirmed through quantitative analyses
362 (Figures 6B and S10B).

363 **Deletion of *MALAT1* Intron 2 using CRISPR/Cas9 impairs cell migration**

364 *MALAT1* has been shown to promote cell migration and metastasis (34, 56) and its expression has
365 been correlated with high tumor progression and metastasis in various tumors (57). We therefore
366 tested whether the intronic sequences that regulate *MALAT1* nuclear speckle localization are
367 related to the ability of *MALAT1* to promote cell migration. To investigate this, we used the

368 metastatic breast cancer cell line MDA-MB-231 and first performed RNA-FISH for *MALAT1* after
369 knocking down *U2AF2*. Similar to SKHEP1 and HCT116 cells, *U2AF2* knockdown resulted in
370 decreased localization of *MALAT1* to speckles in MDA-MB-231 cells (Figures 7A and 7B). We
371 next utilized CRISPR/Cas9 to generate KO clones in MDA-MB-231 cells where either the whole
372 *MALAT1* locus or intron 2 of *MALAT1* were deleted. *MALAT1* depletion in the *MALAT1* KO clone
373 was confirmed with RT-qPCR (Figure S11A). RNA-FISH in 2 separate clones showed that intron
374 2 deletion reduced *MALAT1* localization to nuclear speckles compared to the WT clone (Figures
375 7C-7E). In the intron-2 deletion clones, the transcripts containing intron 2 were, as expected,
376 undetected with primers specifically amplifying intron 2. Interestingly, there was a modest
377 increase in total *MALAT1* expression suggesting a compensatory mechanism for the loss of intron
378 2 (Figure S11B). Functionally, when we tested the migration potential of these cells in transwell
379 cell migration assays, we found significantly reduced cell migration in the intron 2-deleted clones,
380 phenocopying the complete loss of *MALAT1* locus (Figures 7F and S11C). These data suggest that
381 the retention of intron 2 is essential for *MALAT1* localization to nuclear speckles and it plays a key
382 role in promoting cell migration. Together, our data uncovers a U2AF2-driven intron retention
383 program in lncRNAs that regulates cell proliferation and migration.

384 **Discussion**

385 The goal of this study was to examine the causes and consequences of IR in lncRNAs. Intron
386 retention is a regulated process that shapes gene expression networks through mechanisms such as
387 NMD and influences alternative protein isoform generation during differentiation, development,
388 and disease (17, 58). We previously identified a p53-induced lncRNA, *PURPL*, which exhibits
389 growth-promoting functions, and that intron 2 of *PURPL* is retained in multiple cell lines (45). To
390 accomplish our goal, we employed an innovative RNA-seq coupled genome-wide CRISPR screen

391 on a reporter containing *PURPL* intron 2, and unexpectedly identified U2AF2 as a major inhibitor
392 of intron 2 removal in *PURPL*. U2AF2 forms a heterodimer with U2AF1, binding to the Py-tract
393 and the 3′ ss, respectively, to promote splicing by recruiting core spliceosomal components (8, 9,
394 59). Therefore, given the established role of U2AF2 as a critical positive regulator of splicing
395 catalysis, we were surprised that U2AF2 inhibits the removal of *PURPL* intron 2. While U2AF1
396 was not identified as a hit in our screen, follow-up experiments revealed that U2AF1 also
397 suppresses *PURPL* intron 2 removal, suggesting a previously unrecognized role for the U2AF
398 complex as a negative regulator *PURPL* IR.

399 Beyond IR in *PURPL*, we sought to determine whether U2AF2 promotes IR in other
400 transcripts. Using IR Finder (55), we identified other genes where U2AF2 promoted IR. One of
401 the genes that drew our attention was the lncRNA *MALAT1* which showed two intron retention
402 events regulated and bound by U2AF2. These introns contain weak splice sites suggesting that this
403 may be the reason that they are regulated, and they are not constitutively spliced. Because *MALAT1*
404 interacts with splicing factors within nuclear speckles (38), we examined whether U2AF2 affects
405 *MALAT1* localization to speckles. Indeed, our results establish *MALAT1* intron 2 as a critical
406 determinant of its nuclear speckle localization. Future studies involving mapping of the intronic
407 sequences and targeted mutagenesis of the splice sites within *MALAT1* intron 2 could help
408 distinguish whether speckle localization of *MALAT1* is inhibited by the splicing process itself or
409 depends on a discrete element within the intron.

410 For *PURPL*, previous work established its context-dependent oncogenic activity (44, 45),
411 but those studies depleted all transcripts from the *PURPL* locus despite the presence of multiple
412 alternatively spliced isoforms. In the current study, overexpression of the intron 2-retaining
413 *PURPL* isoform in CRISPRi-mediated *PURPL*-depleted cells increased proliferation,

414 demonstrating isoform-specific, growth-promoting function. Our findings reveal that intron
415 retention in both *MALAT1* and *PURPL* is a regulatory mechanism with direct consequences on
416 lncRNA localization and cancer-related phenotypes.

417 The surprising finding of the non-canonical function of U2AF2 promoted us to determine
418 the underlying mechanism. We first asked whether U2AF2 influences transcription initiation or
419 elongation at the *PURPL* locus, given the tight coupling between transcription and splicing (60,
420 61). Transcription kinetics can modulate IR (15, 62) and lncRNAs often lack proximal RNA-pol
421 II phosphorylation over 5′ ss (33). We therefore assayed for PolII-pSer5 (transcription initiation)
422 and PolII-pSer2 (transcription elongation) occupancy after U2AF2 knockdown using CUT&RUN-
423 Seq but there was no significant difference around intron 2 or elsewhere across the *PURPL* locus
424 (data not shown). We also examined the chromatin mark H3K36me3 which has been linked to IR
425 (63) and again observed no effect after U2AF2 depletion (data not shown).

426 U2AF2 functions by binding to the Py-tract, and the Py-tract of *PURPL* intron 2 and
427 *MALAT1* introns 1 and 2 deviate from the consensus, containing a mix of pyrimidines and purines.
428 These atypical sequences likely contribute to splicing inhibition as revealed by our data from the
429 mini-gene reporter for *PURPL*. Given that U2AF1 and U2AF2 bind to multiple intronic sites, a
430 plausible model is that their docking obstructs subsequent spliceosomal assembly. lncRNAs
431 generally undergo less efficient splicing and consequently, they have higher levels of IR compared
432 to protein coding genes (31, 33). A study showed that lncRNA splicing strength correlates with 5′
433 ss strength and lncRNA Py-tract has higher thymidine content compared to protein coding genes
434 (64). The reduced evolutionary pressure to maintain efficient splicing as they are not used as
435 templates for protein synthesis, likely contributes to this pattern. This is also supported by their
436 predominant nuclear localization (26, 27).

437 Our data shows that *PURPL* transcripts can be either nuclear or cytoplasmic, and U2AF2
438 regulation of IR offers a mechanism of nuclear detention. Similar to this notion, the lncRNA *TUG1*
439 remains in the nucleus due to intron retention (65), and the mouse lncRNA *pCharme* undergoes
440 an IR event which renders it chromatin-bound through its association with MATR3 and PTBP1
441 (66). Another aspect of the effects of IR lies in the turnover of the transcripts where additional
442 sequences added by the introns may cause lower transcript stability. In support of this, nuclear
443 lncRNAs have been shown to be less stable than cytoplasmic ones (28), and *PURPL* also follows
444 this rule. Notably, the human lncRNA *hFAST* is cytoplasmic and promotes pluripotency through
445 WNT signaling, but the mouse homolog does not because of its nuclear localization (67).

446 Introns have been proposed to act as docking elements that tether lncRNAs to chromatin
447 (27, 66), and a recent study showed that nuclear speckles are associated with distinct classes of
448 retained introns (68). This raises the question of what makes intron 2 important for speckle
449 localization of *MALAT1*. One possibility is that this region recruits specific RBPs or adopts a
450 structural configuration required for speckle association. Our data further indicates that intron 2
451 retention contributes to *MALAT1's* role in cell migration, potentially because speckle localization
452 is necessary to sustain this phenotype. Future studies are warranted to dig deeper into the
453 underlying mechanism and to determine how intron retention shapes the function of *PURPL* and
454 *MALAT1*.

455 Collectively, our findings reveal an expected function of U2AF2 in promoting intron
456 retention that governs the subcellular localization and function of the lncRNAs *PURPL* and
457 *MALAT1.* As intron retention continues to emerge as a regulated form of alternative splicing, the
458 evidence presented here expands the framework for the mechanisms and biological consequences
459 of intron retention.

460 **Materials and Methods**

461 **Cell culture and treatment**

462 HAP1, HCT116, HEK293T, HepG2, RPE1, SKHEP1, U2OS, WI38, and MDA-MB-231 cells
463 were purchased from ATCC and maintained in Dulbecco's Modified Eagle Medium (DMEM)
464 (Gibco) supplemented with 10% (v/v) fetal bovine serum (Gibco) and 1% (v/v) PenStrep (Gibco)
465 in a 5% CO_2 atmosphere at 37°C. All cell lines were routinely confirmed to be free of mycoplasma
466 using Venor GeM Mycoplasma Detection Kit (Millipore Sigma-Aldrich).

467 **CRASP-Seq *PURPL* screen**

468 The *PURPL* minigene reporter, composed of exons 2 and 3 with the full native intronic sequence,
469 was cloned into the CRASP-Seq genome-wide library as described previously (46). Reporter
470 fragments were synthesized by TWIST Biosciences and assembled into the Eco32I-digested
471 CRASP-Seq knockout library using NEBuilder (49). Library coverage (>250-fold) was maintained
472 through parallel cloning reactions and electroporation into Endura competent cells followed by
473 plasmid amplification and extraction.

474 Lentivirus was produced following established protocols (49). The pooled library was used
475 to infect HAP1 and RPE1 cells expressing *Sp*Cas9 and opCas12a nucleases (49). at a multiplicity
476 of infection (MOI) of ~0.2. After puromycin selection (2 µg/mL for 48 hours), cells were pooled,
477 induced with doxycycline to activate reporter expression and harvested 24 hours later for RNA
478 analysis.

479 Total RNA was extracted, poly(A)+ RNA isolated, and cDNA synthesized using a U6-
480 annealing primer. Spliced and intron-retaining reporter transcripts were amplified using exon- or
481 intron-specific primers incorporating Illumina barcodes. Libraries were sequenced on the NovaSeq
482 6000 platform with a 25% PhiX spike-in and a 300-cycle kit. The sequencing strategy consisted

483 of the following configuration: Read 1 (210 cycles), Index Read 1 (8 cycles), Index Read 2 (8
484 cycles) and Read 2 (85 cycles). Each sample yielded over 500 million total reads.

485 **CRASP-seq data analysis**

486 CRASP-Seq data were processed using a custom pipeline (46). Guides, cell barcodes, and UMIs
487 were retrieved using STAR, and alignment were managed with SAMTools. Reads containing
488 complete hybrid guide, barcodes, UMI, and splicing outcomes were deduplicated prior to analysis.
489 Splicing outcomes were quantified as percent intron retention (PIR): $PIR = [IE / (IE + EE)] \times 100$,
490 where IE represents intron retention reads, and EE represents exon-exon junction reads.
491 Differential PIR (ΔPIR) was calculated as: $\Delta PIR = PIR_{guide} - PIR_{intergenic}$

492 The ΔPIR for each gene and for each guide separately is shown in Table S1. Guides with
493 fewer than 10 reads were excluded. A ΔPIR threshold corresponding to 10% FDR (based on
494 intergenic controls) defined significant hits. At the gene level, hits required FPKM ≥ 0.1, at least
495 two qualifying guides, and $\geq$50% of guides exceeding the ΔPIR threshold in both technical
496 replicates. Genes identified through combinatorial targeting that met these criteria were included
497 as hits (Table S1). The heatmap and Venn diagram in Figure 1C and 1E display single-gene hits
498 as well as collapsed gene hits resulting from combinatorial targeting.

499 **eCLIP-seq data and functional enrichment analysis**

500 eCLIP-seq data for U2AF2, U2AF1, PTBP1, PRPF8, and SRSF1 were obtained from the
501 ENCODE database (encodeproject.org) with the following identifiers: ENCFF989JBA (CTRL),
502 ENCFF566CFJ (CTRL), ENCFF536AFD (U2AF2), ENCFF913WRH (U2AF2), ENCFF159SPZ
503 (U2AF2), ENCFF368XEI (U2AF2), ENCFF542NOA (U2AF1), ENCFF651VRQ (U2AF1),
504 ENCFF363UDO (PTBP1), ENCFF044RKN (PRPF8), and ENCFF327NVE (SRSF1). For
505 visualization of RNA-seq data, the Integrated Genome Viewer (IGV) version 17 was used. The

506 functional enrichment analysis was performed using g:Profiler (69) and the results are shown in
507 the 'gProfiler_Results' tab in Table S1.

508 **siRNA transfections**

509 For siRNA transfections, we used AllStars Negative Control siRNA (QIAGEN). U2AF2 siRNAs
510 were purchased from Dharmacon (J-012380-18 and J-012380-19). U2AF1 siRNA was also
511 purchased from Dharmacon (J-012325-10). Cells were reverse transfected with siRNA at a final
512 concentration of 20 nM using Lipofectamine RNAiMAX (Invitrogen) in Opti-MEM I Reduced
513 Serum Medium (Gibco) according to the manufacturer's protocol. SKHEP1 cells were transfected
514 for 72 hr before RNA harvesting. U2OS, HepG2 and WI38 cells were transfected for two rounds
515 of 48 h + 72 h or 72 h + 72 h. For RNA FISH, SKHEP1, HCT116, and MDA-MB-231 cells were
516 transfected for 72 hr. SKHEP1 expressing the *PURPL* Py-tract mutants were transfected for two
517 rounds of 48 h and 24 h before harvesting, the cells were treated with 1 μg/mL doxycycline to
518 induce the transgene expression. For RNA-seq, the cells were transfected for 72 h before harvested.
519 For RNA stability assay, the cells were reseeded in 12-well plates 48 h after siRNA transfection
520 and 24 h later they were treated with Act D for the indicated time points at 5 μg/mL before RNA
521 extraction and subsequent assays. For RNA-IP, the cells were transfected for 72 h before harvested
522 in 10 cm-plates.

523 **RNA extraction, RT-qPCR, semi-quantitative RT-PCR, and splicing gels**

524 Total RNA was extracted using TRIzol (Invitrogen) according to the manufacturer's protocol.
525 cDNA was then made using iScript™ Reverse Transcription Supermix (Bio-Rad). For real-time
526 qPCR, all reactions were carried out on StepOnePlus real-time PCR System (Applied Biosystems)
527 using FastStart SYBR Green Master Mix (Millipore Sigma). *GAPDH* mRNA was used to
528 normalize expression except for RNA-IPs in Figure 2B where *18S* was used for normalization, and

529 the relative expression/enrichment of RNAs was calculated using the $2^{-\Delta\Delta Ct}$ method. RT-qPCR
530 primer sequences for each gene are indicated in Table S3. For splicing gels, cDNA was used to
531 amplify the region flanking the retained intron with primers in the flanking exons with the addition
532 of one primer inside the intron as indicated in Figure S1B. In the case of *MALAT1*, primers in the
533 flanking exons were utilized due to the short length of introns. For the assay of the effect of U2AF2
534 on the Py-tract mutants of *PURPL*, primers on the transgene were used instead of the endogenous
535 *PURPL* exon 2 and 3. The sequences of the primers can be found in Table S5.

536 **RNA-IPs**

537 For RNA-IP, two 10 cm-plates of SKHEP1 cells were lysed with RIPA buffer (Thermo Scientific)
538 for 10 min on ice before spin down for 10 min at 4ºC. The lysate was incubated with 1 µg of anti-
539 U2AF2 (CST) or rabbit IgG (CST) in 500 µL total volume overnight at 4ºC. The next day, 50 µL
540 of Protein A/G Dynabeads (Pierce) were added to the mix and incubated for 2 h at 4ºC. Beads
541 were then washed 4 times with RIPA buffer and RNA was isolated using TRIzol. Equal volumes
542 of extracted RNA were used for cDNA synthesis and RT-qPCR as described above.

543 **Immunoblotting**

544 For immunoblotting, total cell lysate was prepared using RIPA buffer (Thermo Scientific)
545 containing protease inhibitor cocktail (Roche). Protein concentration was determined using the
546 BCA protein quantitation kit (Thermo Scientific). Cell lysate was loaded onto SDS-PAGE gel and
547 transferred to a PVDF membrane (Thermo Scientific) using a semi-dry apparatus (Bio-Rad). The
548 membrane was blocked in 5% skim milk (Millipore) in TBS containing 0.05% Tween 20. The
549 following antibodies were used: anti-U2AF2 (1:1000, rabbit) (CST, 70471S), anti-GAPDH
550 (1:10,000, rabbit) (CST, 2118L) and secondary anti-rabbit (1:5000) (CST, 7074S).

551 **Nuclear cytoplasmic fractionation**

552 For nuclear cytoplasmic fractionation, SKHEP1 cells were harvested after trypsinization at 80-
553 90% confluence. Cells were pelleted and resuspended in 500 µL RSB buffer (10 mM Tris-HCl,
554 100 mM NaCl, and 2.5 mM MgCl2) containing 40 µg/mL digitonin (Invitrogen), and 100 U of
555 RNase Out. After incubation on ice for 10 min, cell viability was tested with trypan blue to make
556 sure that the cells were permeabilized. The cytoplasmic fraction was harvested as supernatant after
557 centrifugation at 1,500 × g for 5 min at 4°C. The nuclear pellet was resuspended in 100 µL RSB
558 buffer. 100 µL of the cytoplasmic fractionation were transferred to a fresh tube and 900 µL of
559 TRIzol was added as well as to the nuclear extracts for RNA extraction and subsequent cDNA
560 synthesis and RT-qPCR.

561 **Cell proliferation assay**

562 The cell proliferation assay was conducted as previously described (70). To measure proliferation,
563 parental SKHEP1 cells and *PURPL*-CRISPRi cells overexpressing *PURPL* containing intron 2
564 were seeded at 1x10^5 cells per well in 6-well plates in the presence of 1 µg/mL doxycycline. After
565 3 days, the cell numbers and viability were measured. The cells were then passaged at $1x10^5$ and
566 measured again at 6 days, and cumulative cell numbers were plotted.

567 **RNA-seq**

568 The TruSeq Stranded mRNA libraries were pooled and sequenced on one NovaSeq6000 S1
569 flowcell using 2x101 cycles for pair-end run. The Real Time Analysis software (RTA v3.4.4) was
570 used for processing raw data files, the Illumina bcl2fastq v2.20 was used to demultiplex and
571 convert binary base calls and quality scores to fastq format. The sequencing reads were trimmed
572 of adapters and low-quality bases using Cutadapt (v1.18). The trimmed reads were mapped to
573 human reference genome (hg38) and Gencode annotation GENCODE_v30 GTF using STAR
574 aligner (version 2.7.0f) with two-pass alignment option. RSEM (v1.3.1) was used for gene and

575 transcript quantification based on GENCODE annotation file. Palantir Foundry was used for
576 analysis of Differential Expression of Genes inside the secure NIH Integrated Data Analysis
577 Platform (NIDAP). The DEG Analysis results are shown in Table S3.

578 **Iso-Seq**

579 The libraries were prepared using PacBio standard Iso-seq library prep using SMRTbell prep kit
580 3.0. The libraries were sequenced on PacBio Sequel II platform using PacBio v2.0 chemistry.
581 PacBio SMRTlink raw subreads were converted into HiFi circular consensus sequences (CCS)
582 with minimum 3 passes. PacBio IsoSeq v3 pipeline was used to demultiplex the barcodes and
583 remove primers. Additional refine steps included trimming poly-A tails and removing concatemers
584 to generate Full Length Non-Concatemer (FLNC) reads. The FLNC reads were used to map to
585 human reference (hg38) using Minimap2 software to generate alignment bam files. Hierarchical
586 clustering and merging steps were performed to obtain consensus isoforms and the full-length (FL)
587 consensus sequences. The high-quality FL transcripts were mapped to the reference genome by
588 using the minimap2 software. Isoform classification and quality control were done using
589 SQANTI3 software and Illumina short-read gene expression data was combined with PacBio Iso-
590 seq for transcript quantification. The Iso-Seq Analysis results are shown in Table S4. The resulting
591 bam files from the Iso-seq analysis were used for plotting and visualization with the R package
592 Gviz (https://bioconductor.org/packages/release/bioc/html/Gviz.html).

593 **IRFinder**

594 IRFinder (55) was used to identify intron retention events that changed significantly on U2AF2
595 knockdown. A docker image was downloaded from Docker Hub (docker://cloxd/irfinder:2.0.1).
596 The reference was built using pre-downloaded resources (IRFinder BuildRefProcess), using
597 Ensembl release 92 and the 100 bp mappability file provided by IRFinder. The Build-BED-refs.sh

598 script was adjusted so that transcript biotype was not restricted to "processed_transcript" or
599 "protein_coding". Intron retention was quantified using IRFinder in FastQ mode with adapter
600 trimming. Differential intron retention events between siU2AF2 vs. siCTRL were identified with
601 IRFinder Diff using DESeq2 with the default IR ratio ≥ 0.05 in at least one sample and no warning
602 filters, which profiled 16,859 unique introns.

603 **Construct generation**

604 For Py-tract mutants, gene fragments were purchased from Twist Biosciences and cloned into the
605 pLCHKO-puro vector containing the WT sequences of *PURPL* (Figure 1A), between sites for
606 NseI and AfeI. The *PURPL* transcript containing intron 2 was cloned into the pCW57 vector
607 (Hygro) (Addgene #80922) in SalI and MluI sites. *MALAT1* transcripts were cloned into the
608 pCW57 vector (BSD) (Addgene #80921) in SalI and MluI sites. The *PURPL* insert was generated
609 by Twist Biosciences and the *MALAT1* sequence was amplified using Phusion DNA polymerase
610 (NEB). For deletion mutants, we used the Q5® Site-Directed Mutagenesis Kit (NEB). The primers
611 used for cloning are listed in Table S5.

612 **Generation of *MALAT1* KO cell lines**

613 For HCT116 *MALAT1* knockout, we used integration by non-homologous end joining, which was
614 accomplished by introducing a simultaneous double-strand break in genomic DNA and in the
615 targeting vectors (pCMV-Puro and pCMV-Hygro) (71, 72). Plasmids encoding spCas9 and
616 sgRNAs were obtained from Addgene (Plasmids #41815 and #47108). Oligonucleotides for
617 construction of sgRNAs were obtained from Integrated DNA Technologies, hybridized,
618 phosphorylated and cloned into the sgRNA plasmid or targeting vector using BbsI sites (73).
619 Target sequences for sgRNAs are provided in Table S5. These vectors were transfected into

620 HCT116 using Lipofectamine 2000 following recommendations from the manufacturer. Three
621 days after transfection, the cells were selected with Puromycin (0.5 μg/mL) and Hygromycin (100
622 μg/mL) to generate clonal populations. Genomic DNA from each clone was isolated using DNEasy
623 Blood and Tissue Kit (Qiagen). PCRs to detect integration of the targeting vector at the target site
624 were performed using KAPA2G Robust PCR kits (Kapa Biosystems) according to the
625 manufacturer's instructions. A typical reaction contained 20-100 ng of genomic DNA in Buffer A
626 (5 μL), Enhancer (5 μL), dNTPs (0.5 μL), primers forward (PINCR Det FP, 1.25 μL) and reverse
627 (Targeting vector Det RP, 1.25 μL) and KAPA2G Robust DNA Polymerase (0.5 U). The DNA
628 sequences of the primers for each target are provided in Table S5. PCR products were visualized
629 in 2% agarose gels and images were captured using a ChemiDoc-It2 (UVP).

630 To generate MDA-MB-231 whole-locus and intron 2 *MALAT1* knockout clones, we used
631 the CRIPSR/Cas9 system by transfecting the cells with gRNAs and purified Cas9. 2 μL of each
632 30 μM gRNA was mixed with 1 μL (~5 μg) of purified Cas9 and 100 μL of 4D-Nucleofector® X
633 SE buffer SE from the Cell Line Kit (Lonza) supplemented with 18.5 μL of SE supplement per
634 100 μL total buffer volume. The mixture was incubated at room temperature for 10 min and
635 1x10^6 cells were resuspended in nucleofection cuvettes. Nucleofections were carried out using
636 the 4D-Nucleofector® System (Lonza). The cells were transferred to a 6-well plate, and they were
637 single-cell sorted 48 h later. The clones were grown for 3-4 weeks before genomic DNA extraction
638 and genotyping with PCR using the primers in Table S5.

639 ***PURPL* and *MALAT1* overexpression and transduction**

640 *PURPL* intron 2 constructs were transduced in previously generated *PURPL*-CRISPRi SKHEP1
641 cells (45). *MALAT1*-WT and deletion mutants were transduced in the two HCT116-*MALAT1* KO
642 clones. For lentivirus generation, HEK293T cells were seeded at 2×10^5 cells/well onto 6-well

643 plates, and 1,200 ng of each construct was transfected with lentiviral package vectors using
644 Lipofectamine 2000 (Life Technologies, Invitrogen) according to the manufacturer's protocol.
645 Virus-containing media were harvested at 48 and 72 h post-transfection and added to cells for 24
646 h. Transduced cells were selected with 500 μg/mL hygromycin for 2 weeks (SKHEP1) or 10
647 μg/mL Blasticidin (HCT116).

648 **RNA-FISH, IF and Quantification of FISH**

649 The *PURPL* intron probe set was custom-designed using the Stellaris Probe Designer, tagged with
650 Quasar 570 dye and purchased from Stellaris. The *MALAT1* smFISH probe set with Quasar 670
651 dye was purchased from Stellaris (Cat# VSMF-2211-5). For *PURPL* intron smRNA-FISH,
652 SKHEP1 cells were seeded on coverslips and cultured for 24 hours at 37°C with 5% CO_2 before
653 inducing DNA damage by incubating the cells with 2 mM hydroxyurea for 24 hours (or regular
654 growth media for control). For *MALAT1* smRNA-FISH, HCT116 and MDA-MB-231 *MALAT1*
655 KO cells were seeded on coverslips and cultured for 24 hours at 37°C with 5% CO_2. *MALAT1* WT,
656 *MALAT1* del-1, *MALAT1* del-2, *MALAT1* del1+2, or empty vector expression was induced by
657 incubating the cells with 1 μg/mL doxycycline for 48 hours. The cells that were transfected with
658 siRNA for U2AF2 were seeded on coverslips at the time of transfection. Cells were fixed with 4%
659 PFA for 15 min at room temperature, permeabilized with 0.5% Triton X-100, and washed with
660 washing buffer (10% formamide, 2XSSC) for 5 min. Probe was added to hybridization buffer
661 (10% formamide in Stellaris hybridization buffer Cat# SMF-HB1-10) at a final concentration of
662 125 nM. Hybridization was done in a humidified chamber in the dark overnight at 37°C. After
663 hybridization, the coverslips were washed twice with wash buffer for 30 min at 37°C and then
664 post-fixed with 4% PFA for 10 min at room temperature. The coverslips were blocked with 5%
665 normal goat serum and then incubated with SON antibody (Sigma Cat# HPA023535). Primary

666 and secondary antibodies were diluted in 1% normal goat serum. Antibody incubations were done
667 at room temperature for 1 h and washed with PBS. DNA was counterstained by DAPI. The
668 coverslips were then washed with 4XSSC for 5 min at room temperature and mounted in
669 VectaShield Antifade Mounting Medium (Vector Laboratories). Images were taken using
670 DeltaVision microscope (GE Healthcare) equipped with 60 X/1.42 NA oil immersion objective
671 (Olympus) and CoolSNAP-HQ2 camera. To quantify the speckle to nucleoplasm signal intensity
672 ratio of *MALAT1* in Figures 5B, 7B, 7D, and S8B ImageJ was used to create nucleus masks based
673 on DAPI and speckles mask based on SON. The signal intensity of *MALAT1* in the speckles mask
674 was divided by the *MALAT1* signal intensity in the nucleus-minus-speckles area. For the
675 quantification of Figures 6B, 7E and S10B, the cells with *MALAT1* FISH signal showing
676 prominent speckle patterns similar to SON were classified as "Speckle" pattern, and the cells with
677 a diffused *MALAT1* FISH signal were classified as "Diffused". Only *MALAT1*-positive cells were
678 counted for quantification.

679 **Transwell cell migration assays**

680 For the transwell assay, WT and KO MDA-MB-231 clones were seeded at $10x10^5$ cells per well
681 in 12-well plates. 24 h later, the media were removed and replaced with media containing 1% FBS.
682 After 5 h, the cells were trypsinized and seeded in the top chamber of the transwell plates (Corning,
683 Catalog#354578) at 25,000 cells per well at a volume of 100 µL. The lower chamber of the
684 transwell is filled with 700 µL of media containing 10% FBS. The cells were incubated for 24 h
685 and the transwell were removed from the plates. The non-migrated cells were scraped with a cotton
686 swab, and the migrated cells were fixed with ice-cold methanol for 5 min and stained with 0.05%
687 crystal violet blue. The migrated cells were counted under a light microscope.

688 **Statistics**

689 Statistics were performed using the Student's t-test.

690 **Data availability**

691 The Iso-Seq data files have been deposited at GEO with accession number GSE288684 and
692 reviewer token: ktcnasayrbgnjev (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi). The RNA-
693 Seq data files have been deposited at GEO with accession number GSE288685 and reviewer token:
694 qvwhuawixxqjnml (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi).

695 **Acknowledgments**

696 This research was supported by the Intramural Research Program (A.L.) of the National Cancer
697 Institute (NCI), Center for Cancer Research (CCR) (project numbers ZIA BC011646, ZIA
698 BC012019), NIH. KVP laboratory is supported by grants from NIH (R01-GM132458), Cancer
699 center at Illinois seed grant, ARPA-H, and National Science Foundation (NSF) center for
700 quantitative cell biology. YJS is supported by NIH T32EB019944. We thank the CCR Genomics
701 Core, CCR, NCI, Bethesda, MD for its valuable assistance with Sanger and Nanopore sequencing.
702 We also thank the Sequencing Facility, CCR, NCI, Frederick, MD for performing the RNA
703 sequencing assays. The contributions of the NIH author(s) were made as part of their official duties
704 as NIH federal employees, are in compliance with agency policy requirements, and are considered
705 Works of the United States Government. However, the findings and conclusions presented in this
706 paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S.
707 Department of Health and Human Services.

708 **Author Contributions**

709 IG, CN, YJS, CCH, SK, RP, XLL, RK, RS, and PPZ conducted experiments. IG, AKB, EP, YZ,
710 BS, TB, XW, TGP, KP, and AL did data analysis. IG and AL wrote the manuscript. NC, KP, and
711 TGP edited the manuscript.

712 **Declaration of interests**

713 The authors declare they have no competing interest.

714 **References**

715 1. M. E. Wilkinson, C. Charenton, K. Nagai, RNA Splicing by the Spliceosome. *Annu Rev*
716 *Biochem* **89**, 359-388 (2020).
717 2. L. E. Marasco, A. R. Kornblihtt, The physiology of alternative splicing. *Nat Rev Mol Cell*
718 *Biol* **24**, 242-254 (2023).
719 3. A. G. Matera, Z. Wang, A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* **15**,
720 108-121 (2014).
721 4. D. L. Black, Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*
722 **72**, 291-336 (2003).
723 5. B. Ruskin, P. D. Zamore, M. R. Green, A factor, U2AF, is required for U2 snRNP
724 binding and splicing complex assembly. *Cell* **52**, 207-219 (1988).
725 6. J. A. Berglund, K. Chua, N. Abovich, R. Reed, M. Rosbash, The splicing factor BBP
726 interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* **89**,
727 781-787 (1997).
728 7. J. A. Berglund, N. Abovich, M. Rosbash, A cooperative interaction between U2AF65 and
729 mBBP/SF1 facilitates branchpoint region recognition. *Genes Dev* **12**, 858-867 (1998).
730 8. S. Wu, C. M. Romfo, T. W. Nilsen, M. R. Green, Functional recognition of the 3' splice
731 site AG by the splicing factor U2AF35. *Nature* **402**, 832-835 (1999).
732 9. D. A. Zorio, T. Blumenthal, Both subunits of U2AF recognize the 3' splice site in
733 Caenorhabditis elegans. *Nature* **402**, 835-838 (1999).

734 10. E. T. Wang *et al.*, Alternative isoform regulation in human tissue transcriptomes. *Nature*
735 **456**, 470-476 (2008).
736 11. T. W. Nilsen, B. R. Graveley, Expansion of the eukaryotic proteome by alternative
737 splicing. *Nature* **463**, 457-463 (2010).
738 12. Q. Pan, O. Shai, L. J. Lee, B. J. Frey, B. J. Blencowe, Deep surveying of alternative
739 splicing complexity in the human transcriptome by high-throughput sequencing. *Nat*
740 *Genet* **40**, 1413-1415 (2008).
741 13. R. A. Gier *et al.*, High-performance CRISPR-Cas12a genome editing for combinatorial
742 genetic screening. *Nat Commun* **11**, 3455 (2020).
743 14. H. Keren, G. Lev-Maor, G. Ast, Alternative splicing and evolution: diversification, exon
744 definition and function. *Nat Rev Genet* **11**, 345-355 (2010).
745 15. U. Braunschweig *et al.*, Widespread intron retention in mammals functionally tunes
746 transcriptomes. *Genome Res* **24**, 1774-1786 (2014).
747 16. J. J. Wong *et al.*, Orchestrated intron retention regulates normal granulocyte
748 differentiation. *Cell* **154**, 583-595 (2013).
749 17. G. Monteuuis, J. J. L. Wong, C. G. Bailey, U. Schmitz, J. E. J. Rasko, The changing
750 paradigm of intron retention: regulation, ramifications and recipes. *Nucleic Acids Res* **47**,
751 11497-11513 (2019).
752 18. T. Eom *et al.*, NOVA-dependent regulation of cryptic NMD exons controls synaptic
753 protein levels after seizure. *Elife* **2**, e00178 (2013).
754 19. M. A. Hossain *et al.*, Posttranscriptional Regulation of Gcr1 Expression and Activity Is
755 Crucial for Metabolic Adjustment in Response to Glucose Availability. *Mol Cell* **62**, 346-
756 358 (2016).

757 20. S. Cha, C. P. Hong, H. A. Kang, J. S. Hahn, Differential activation mechanisms of two
758 isoforms of Gcr1 transcription factor generated from spliced and un-spliced transcripts in
759 Saccharomyces cerevisiae. *Nucleic Acids Res* **49**, 745-759 (2021).
760 21. U. Schmitz *et al.*, Intron retention enhances gene regulatory complexity in vertebrates.
761 *Genome Biol* **18**, 216 (2017).
762 22. H. Dvinge, R. K. Bradley, Widespread intron retention diversifies most cancer
763 transcriptomes. *Genome Med* **7**, 45 (2015).
764 23. P. L. Boutz, A. Bhutkar, P. A. Sharp, Detained introns are a novel, widespread class of
765 post-transcriptionally spliced introns. *Genes Dev* **29**, 63-80 (2015).
766 24. J. J. Quinn, H. Y. Chang, Unique features of long non-coding RNA biogenesis and
767 function. *Nat Rev Genet* **17**, 47-62 (2016).
768 25. I. Ulitsky, D. P. Bartel, lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26-
769 46 (2013).
770 26. C. J. Guo, G. Xu, L. L. Chen, Mechanisms of Long Noncoding RNA Nuclear Retention.
771 *Trends Biochem Sci* **45**, 947-960 (2020).
772 27. B. Zuckerman, I. Ulitsky, Predictive models of subcellular localization of long RNAs.
773 *RNA* **25**, 557-572 (2019).
774 28. M. B. Clark *et al.*, Genome-wide analysis of long noncoding RNA stability. *Genome Res*
775 **22**, 885-898 (2012).
776 29. M. N. Cabili *et al.*, Localization and abundance analysis of human lncRNAs at single-cell
777 and single-molecule resolution. *Genome Biol* **16**, 20 (2015).

778 30. H. Tilgner *et al.*, Deep sequencing of subcellular RNA fractions shows splicing to be
779 predominantly co-transcriptional in the human genome but inefficient for lncRNAs.
780 *Genome Res* **22**, 1616-1625 (2012).
781 31. M. Mele *et al.*, Chromatin environment, transcriptional regulation, and splicing
782 distinguish lincRNAs and mRNAs. *Genome Res* **27**, 27-37 (2017).
783 32. M. Kumar, I. Grammatikakis, Causes, Consequences and Challenges of Intron Retention
784 in lncRNAs. *Mol Cell Biol* **45**, 576-584 (2025).
785 33. N. Mukherjee *et al.*, Integrative classification of human coding and noncoding genes
786 through RNA metabolism profiles. *Nat Struct Mol Biol* **24**, 86-96 (2017).
787 34. T. Gutschner *et al.*, The noncoding RNA MALAT1 is a critical regulator of the
788 metastasis phenotype of lung cancer cells. *Cancer Res* **73**, 1180-1189 (2013).
789 35. J. N. Hutchinson *et al.*, A screen for nuclear transcripts identifies two linked noncoding
790 RNAs associated with SC35 splicing domains. *BMC Genomics* **8**, 39 (2007).
791 36. P. Ji *et al.*, MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis
792 and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031-8041 (2003).
793 37. D. Bernard *et al.*, A long nuclear-retained non-coding RNA regulates synaptogenesis by
794 modulating gene expression. *EMBO J* **29**, 3082-3093 (2010).
795 38. V. Tripathi *et al.*, The nuclear-retained noncoding RNA MALAT1 regulates alternative
796 splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**, 925-938 (2010).
797 39. M. K. Shinn *et al.*, Nuclear speckle proteins form intrinsic and MALAT1-dependent
798 microphases. *Cell* **189**, 832-852 e824 (2026).
799 40. D. L. Spector, A. I. Lamond, Nuclear speckles. *Cold Spring Harb Perspect Biol* **3** (2011).

800 41. A. Sharma, H. Takata, K. Shibahara, A. Bubulya, P. A. Bubulya, Son is essential for
801 nuclear speckle organization and cell cycle progression. *Mol Biol Cell* **21**, 650-663
802 (2010).
803 42. S. Nakagawa *et al.*, Malat1 is not an essential component of nuclear speckles in mice.
804 *RNA* **18**, 1487-1499 (2012).
805 43. B. Zhang *et al.*, The lncRNA Malat1 is dispensable for mouse development but its
806 transcription plays a cis-regulatory role in the adult. *Cell Rep* **2**, 111-123 (2012).
807 44. X. L. Li *et al.*, Long Noncoding RNA PURPL Suppresses Basal p53 Levels and
808 Promotes Tumorigenicity in Colorectal Cancer. *Cell Rep* **20**, 2408-2423 (2017).
809 45. C. C. R. Hartford *et al.*, Context-Dependent Function of Long Noncoding RNA PURPL
810 in Transcriptome Regulation during p53 Activation. *Mol Cell Biol* **42**, e0028922 (2022).
811 46. A. K. Behera *et al.*, RNA-coupled CRISPR screens reveal ZNF207 as a regulator of
812 LMNA aberrant splicing in progeria. *Mol Cell* **86**, 41-59 e15 (2026).
813 47. M. Aregger, K. Xing, T. Gonatopoulos-Pournatzis, Application of CHyMErA Cas9-
814 Cas12a combinatorial genome-editing platform for genetic interaction mapping and gene
815 fragment deletion screening. *Nat Protoc* **16**, 4722-4765 (2021).
816 48. T. Gonatopoulos-Pournatzis *et al.*, Genetic interaction mapping and exon-resolution
817 functional genomics with a hybrid Cas9-Cas12a platform. *Nat Biotechnol* **38**, 638-648
818 (2020).
819 49. M. S. Xiao *et al.*, Genome-scale exon perturbation screens uncover exons critical for cell
820 fitness. *Mol Cell* **84**, 2553-2572 e2519 (2024).
821 50. E. L. Van Nostrand *et al.*, Robust transcriptome-wide discovery of RNA-binding protein
822 binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514 (2016).

823 51. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome.
824 *Nature* **489**, 57-74 (2012).
825 52. Y. Luo *et al.*, New developments on the Encyclopedia of DNA Elements (ENCODE)
826 data portal. *Nucleic Acids Res* **48**, D882-D889 (2020).
827 53. T. Berhane *et al.*, Knockdown of the long noncoding RNA PURPL induces apoptosis and
828 sensitizes liver cancer cells to doxorubicin. *Sci Rep* **12**, 19502 (2022).
829 54. S. Han *et al.*, PURPL represses autophagic cell death to promote cutaneous melanoma by
830 modulating ULK1 phosphorylation. *Cell Death Dis* **12**, 1070 (2021).
831 55. R. Middleton *et al.*, IRFinder: assessing the impact of intron retention on mammalian
832 gene expression. *Genome Biol* **18**, 51 (2017).
833 56. E. Martinez-Terroba *et al.*, Overexpression of Malat1 drives metastasis through
834 inflammatory reprogramming of the tumor microenvironment. *Sci Immunol* **9**, eadh5462
835 (2024).
836 57. G. Arun, D. Aggarwal, D. L. Spector, MALAT1 Long Non-Coding RNA: Functional
837 Implications. *Noncoding RNA* **6** (2020).
838 58. A. G. Jacob, C. W. J. Smith, Intron retention as a component of regulated gene
839 expression programs. *Hum Genet* **136**, 1043-1057 (2017).
840 59. C. Shao *et al.*, Mechanisms for U2AF to define 3' splice sites and regulate alternative
841 splicing in the human genome. *Nat Struct Mol Biol* **21**, 997-1005 (2014).
842 60. S. Naftelberg, I. E. Schor, G. Ast, A. R. Kornblihtt, Regulation of alternative splicing
843 through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84**, 165-
844 198 (2015).

845 61. L. Herzel, D. S. M. Ottoz, T. Alpert, K. M. Neugebauer, Splicing and transcription touch
846 base: co-transcriptional spliceosome assembly and function. *Nat Rev Mol Cell Biol* **18**,
847 637-650 (2017).
848 62. J. J. Wong *et al.*, Intron retention is regulated by altered MeCP2-mediated splicing factor
849 recruitment. *Nat Commun* **8**, 15134 (2017).
850 63. R. Guo *et al.*, BS69/ZMYND11 reads and connects histone H3.3 lysine 36
851 trimethylation-decorated chromatin to regulated pre-mRNA processing. *Mol Cell* **56**,
852 298-310 (2014).
853 64. Z. Krchnakova *et al.*, Splicing of long non-coding RNAs primarily depends on
854 polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR
855 proteins. *Nucleic Acids Res* **47**, 911-928 (2019).
856 65. G. Dumbovic *et al.*, Nuclear compartmentalization of TERT mRNA and TUG1 lncRNA
857 is driven by intron retention. *Nat Commun* **12**, 3308 (2021).
858 66. F. Desideri *et al.*, Intronic Determinants Coordinate Charme lncRNA Nuclear Activity
859 through the Interaction with MATR3 and PTBP1. *Cell Rep* **33**, 108548 (2020).
860 67. C. J. Guo *et al.*, Distinct Processing of lncRNAs Contributes to Non-conserved Functions
861 in Stem Cells. *Cell* **181**, 621-636 e622 (2020).
862 68. A. R. Barutcu *et al.*, Systematic mapping of nuclear domain-associated transcripts reveals
863 speckles and lamina as hubs of functionally distinct retained introns. *Mol Cell* **82**, 1035-
864 1052 e1039 (2022).
865 69. L. Kolberg *et al.*, g:Profiler-interoperable web service for functional enrichment analysis
866 and gene identifier mapping (2023 update). *Nucleic Acids Res* **51**, W207-W212 (2023).

867 70. J. S. Lee *et al.*, An ultraconserved snoRNA-like element in long noncoding RNA
868 CRNDE promotes ribosome biogenesis and cell proliferation. *Mol Cell* **85**, 1543-1560
869 e1510 (2025).
870 71. A. Brown *et al.*, Multiplexed and tunable transcriptional activation by promoter insertion
871 using nuclease-assisted vector integration. *Nucleic Acids Res* **47**, e67 (2019).
872 72. A. Brown, W. S. Woods, P. Perez-Pinera, Multiplexed Targeted Genome Engineering
873 Using a Universal Nuclease-Assisted Vector Integration System. *ACS Synth Biol* **5**, 582-
874 588 (2016).
875 73. A. Brown, W. S. Woods, P. Perez-Pinera, Targeted Gene Activation Using RNA-Guided
876 Nucleases. *Methods Mol Biol* **1468**, 235-250 (2017).

877

878 **Figure legends**

879 **Figure 1. CRISPR-based screen identifies U2AF2 as a promoter of intron retention. (A)**
880 Schematic of the *PURPL* IR reporter used in the screen to identify IR regulators. The pLCHKO
881 vector contains the puromycin resistance gene used for selection. The transgene also contains the
882 genomic sequences of exon 2 (161 bp), intron 2 (1,943 bp), and exon 3 (163 bp) of the *PURPL*
883 gene under a Doxycycline-inducible promoter. A hybrid guide (hg)RNA for guiding *Sp*Cas9 and
884 *Lb*Cas12a is expressed under a U6 promoter. Top arrow: The hgRNA is cleaved by Cas12a to
885 produce 2 guide RNAs for dual targeting of each protein coding gene, paralog targeting, and
886 genetic interaction mapping. Bottom arrow: Upon intron splicing or intron retention, two different
887 reads are generated. **(B)** Two different cell lines, stably expressing Cas9 and Cas12a, are used:
888 HAP1, and RPE1. The CRASP-seq pipeline until Illumina paired-end sequencing of CRISPR
889 libraries is depicted. **(C)** Heatmap showing single genes only (after removing gene pairs), whether

890 identified individually or as part of a gene pair. The hits are ordered by the average ΔPIR values
891 across HAP1 and RPE1, from lowest to highest. **(D)** Left: Gene ontology enrichment analysis of
892 biological processes regulated by the hits of the screen. The most significant pathways involve
893 mRNA processing and RNA splicing. Right: Gene ontology enrichment analysis of cellular
894 components regulated by the hits of the screen. These components include mainly Spliceosomal
895 Complex, small nuclear ribonucleoprotein complex and the Catalytic step 2 Spliceosome. **(E)**
896 Venn diagram showing that there are 5 common hits which cause a reduction in the percentage of
897 intron retention (ΔPIR) of the splicing reporter in the two cell lines. **(F)** Table showing the top 5
898 hits of the screen with a ΔPIR for each hit in each cell line and their full names.

899 **Figure 2. U2AF2 binds to intron 2-containing *PURPL* transcripts and promotes intron 2**
900 **retention. (A)** IGV snapshot of eCLIP-seq data showing binding sites and enrichment of U2AF2
901 and U2AF1 on the *PURPL* transcripts around intron 2. eCLIP-seq data for PTBP1, PRPF8, and
902 SRSF1 are indicated. The annotated locus by RefSeq is also indicated. eCLIP data-seq was
903 downloaded from encodeproject.org. **(B)** RT-qPCR after RNA-IPs using a U2AF2 antibody with
904 primer pairs specifically detecting *PURPL* transcripts as indicated in Figure S1B. U2AF2 binds to
905 transcripts containing intron 2 but not the ones with intron 1, intron 3, or spliced exons 2 and 3.
906 Samples were normalized to IgG-IP. 18S was used as a loading control. **(C)** Validation of hgRNAs
907 using the *PURPL* minigene in HAP1 and RPE1 cells. The gels show RT-PCR products for *PURPL*
908 transcripts with specific primers for the minigene. The hgRNAs used were targeting intergenic
909 region, or U2AF2. The schematics next to the gel indicate the expected products of the intron-
910 retained and spliced isoforms. **(D)** Western blot for U2AF2 showing successful knockdown of
911 U2AF2 protein in SKHEP1 cells. GAPDH was used as a loading control (lower panel). **(E)**
912 Schematic of the PCR primer triplet used to detect intron 2 retention (red) or splicing (purple). The

913 length for each PCR product is indicated. **(F)** Gel with RT-PCR products for *PURPL* upon
914 knockdown of U2AF2 with 2 different siRNAs in SKHEP1 cells. The schematics next to the gel
915 indicate the expected products of the intron-retained and spliced isoforms. Between the two
916 expected PCR products, we observed an extra band corresponding to the inclusion of an alternative
917 exon inside *PURPL* intron 2 as observed in RefSeq, the inclusion of which is not affected by
918 U2AF2. **(G)** Graph with quantitation of the gel bands in **(F)**. Error bars represent standard
919 deviations from 3 **(G)** experiments. ***p<0.001.

920 **Figure 3. Expression of Intron 2 Retaining *PURPL* leads to high proliferation. (A)** RNA-FISH
921 images for *PURPL* with intron 2 retention and *MALAT1* in HCT116 cells without treatment or
922 after 24 h of 2 mM of Hydorxyurea (HU) to induce *PURPL* expression. Scale bar is 10 μm. **(B)**
923 RNA stability assays were performed for *PURPL* transcripts by measuring their levels by RT-
924 qPCR following 0, 2, and 4 h of ActD treatment in SKHEP1 cells. *MYC* was used as a positive
925 control for unstable RNA. **(C)** RT-qPCR showing *PURPL* depletion in in SKHEP1 cells using 3
926 different gRNAs (sg2, sg4, sg5) compared to a Non-Targeting Control gRNA. **(D)** RT-qPCR for
927 intron 2-containing PURPL transcript after 48 h of 1 μg/mL doxycycline treatment in comparison
928 to no treatment in SKHEP1 *PURPL*-CRISPRi populations using the 3 gRNAs as in **(C)**. **(E)**
929 Proliferation assay showing the effect of overexpression of intron 2-containing *PURPL* transcript
930 in the proliferation of SKHEP1 cells where the endogenous *PURPL* is knocked down with
931 CRISPRi. Error bars represent standard deviation from 3 populations with different gRNAs. The
932 cells were treated with 1 μg/mL doxycycline to induce intron 2-retained *PURPL* expression and
933 cell proliferation was monitored at 3 and 6 days. **(F)** Gel with RT-PCR products for transcripts
934 using primers as indicated in Fig. 2E upon doxycycline treatment of SKHEP1 CRIPSRi cells. The
935 three repeats represent 3 different clones of cells. The last lane is an RT- control. The schematics

936 next to the gel indicate the expected products of the intron-retained and spliced isoforms. NS
937 indicates a non-specific band. Error bars in **(B)**, **(C)**, and **(D)** represent standard deviations from 2
938 experiments. *p<0.05, **p<0.01, ***p<0.001.

939 **Figure 4. U2AF2 directly promotes Intron Retention of *MALAT1*. (A)** U2AF2 was knocked
940 down in SKHEP1 cells and 72 h later, RNA was extracted and RNA-seq was performed. *Left*:
941 Number of decreased (blue) and increased (red) IR events at various p-values after U2AF2
942 knockdown as analyzed with the IR Finder algorithm. The purple arrow indicates the *PURPL* IR
943 event and green arrows indicate *MALAT1* IR events *Right*: Pie chart of the numbers of increased
944 and decreased IR events upon U2AF2 knockdown. **(B)** Floating bar plots showing the IR ratio of
945 *PURPL* intron 2 and the IR ratio of intron 1 (middle) and intron 2 (right) of *MALAT1* in siCTRL
946 and siU2AF2#1 samples as analyzed with the IRFinder algorithm. **(C)** and **(E)** RT-PCR for
947 *MALAT1* using a primer pair flanking the regulated intron 1 **(C)** or intron 2 **(E)** upon knockdown
948 of U2AF2 with 2 different siRNAs in HCT116 and SKHEP1 cells. The schematics next to the gel
949 indicate the expected products of the intron-retained and spliced isoforms. **(D)** and **(F)** Bar graph
950 with quantitation of the gel bands from **(C)** and **(E)** in SKHEP1 cells. Error bars represent standard
951 deviations from 2 independent experiments. *p<0.05, **p<0.01, ***p<0.001.

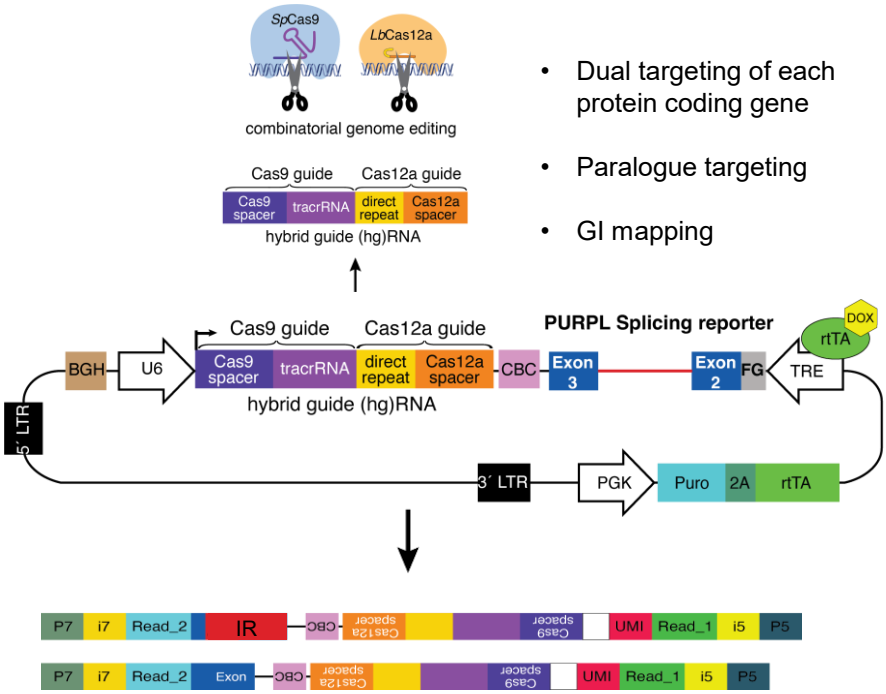
952 **Figure 5. U2AF2 promotes localization of *MALAT1* to nuclear speckles. (A)** RNA-FISH
953 images for *MALAT1* and Immunofluorescence images for SON is shown upon transfection of
954 SKHEP1 cells with siCTRL or siU2AF2 with two different siRNAs. *MALAT1* is enriched in
955 nuclear speckles in the siCTRL but not upon U2AF2 knockdown. **(B)** Quantitation of the speckle
956 to nuclear plasma *MALAT1* signal ratio in the three replicates in panel **(A)**. ****p<0.0001.

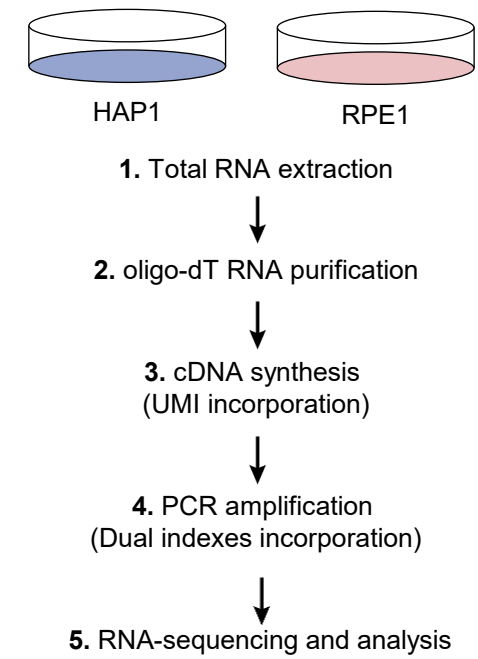
957 **Figure 6. Intron 2 of *MALAT1* dictates its localization to nuclear speckles. (A)** RNA-FISH
958 images for *MALAT1* and Immunofluorescence images for SON in a HCT116 *MALAT1*-KO clone

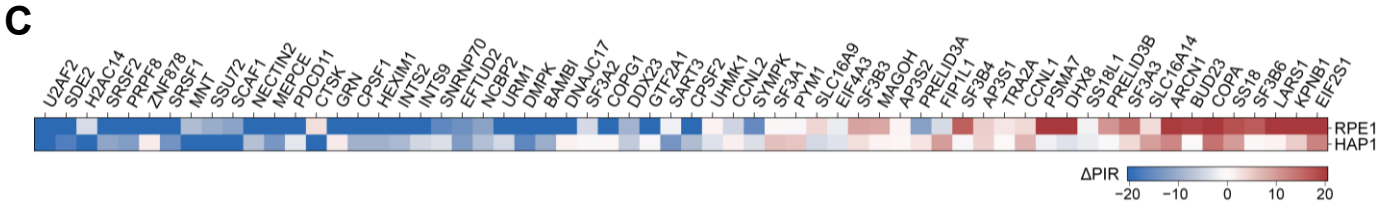
959 (clone 1) where Empty Vector (EV) or constructs that exogenously express *MALAT1* transcripts
960 were re-introduced using a doxy-inducible lentivirus. Expressed *MALAT1* transcripts were full-
961 length (WT) or intron 1 deleted (del-1) or intron 2 deleted (del-2) or both introns deleted
962 simultaneously (del-1+2). The cells were first treated with 1 μg/mL doxycycline for 48 h to induce
963 *MALAT1* expression. *MALAT1* without intron 2 does not enrich in speckles as it is shown with
964 SON protein. Scale bar is 10 μm. **(B)** Graph showing quantification of the percentage of cells
965 where *MALAT1* transcripts are enriched in speckles or being diffused in HCT116 cells clone 1.
966 Error bars represent standard deviations from 2 independent experiments.

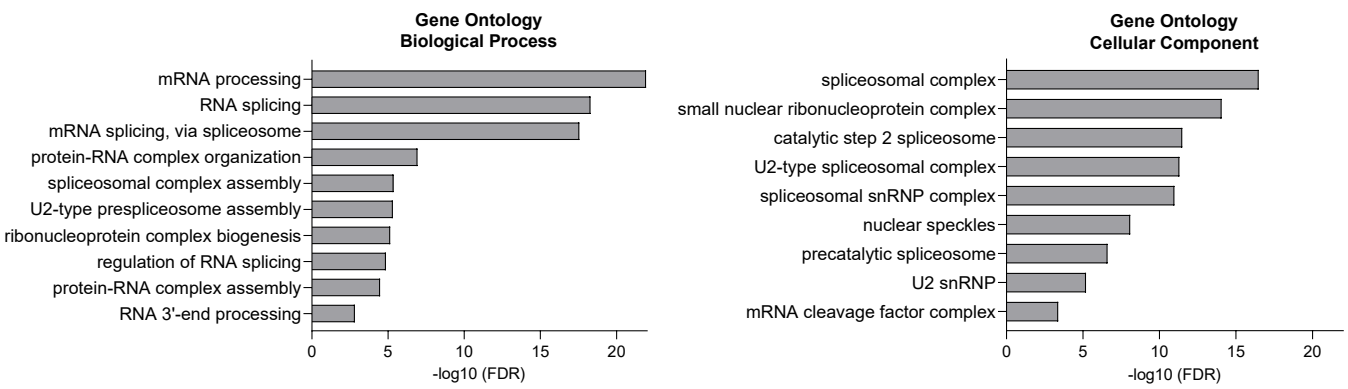
967 **Figure 7. Intron 2 regulates the migration capacity of *MALAT1* in breast cancer cells. (A)**
968 RNA-FISH images for *MALAT1* and Immunofluorescence images for SON are shown upon
969 transfection of MDA-MB-231 cells with siCTRL or siU2AF2. *MALAT1* is enriched in nuclear
970 speckles in the siCTRL but not upon U2AF2 knockdown. **(B)** Quantitation of the speckle to nuclear
971 plasma *MALAT1* signal ratio in the three replicates in panel **(A)**. **(C)** RNA-FISH images for
972 *MALAT1* and Immunofluorescence images for SON in MDA-MB-231 clones either WT, or whole
973 locus *MALAT1* deletion, or intron 2 deletion. When intron 2 is deleted, *MALAT1* does not enrich
974 in speckles as it is shown with SON protein. Scale bar is 10 μm. **(D)** Quantitation of the speckle to
975 nuclear plasma *MALAT1* signal ratio in the three replicates in panel **(C)**. **(E)** Graph showing
976 quantification of the percentage of *MALAT1* transcripts enriched in speckles or being diffused in
977 MDA-MB-231 cells in panel **(C)**, N=2. **(F)** Bar graph showing the number of cells that have
978 migrated in transwell migration assays conducted in MDA-MB-231 WT, *MALAT1* KO and
979 *MALAT1* – intron 2 deletion clones. Intron 2 deletion leads to decreased migration potential. The
980 number of cells are the sum of cells from 5 different fields, N=4. *$p<0.05$, **$p<0.01$, ***$p<0.001$,
981 ****$p<0.0001$.

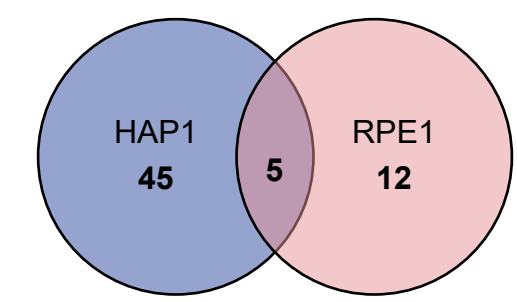
Figure 1







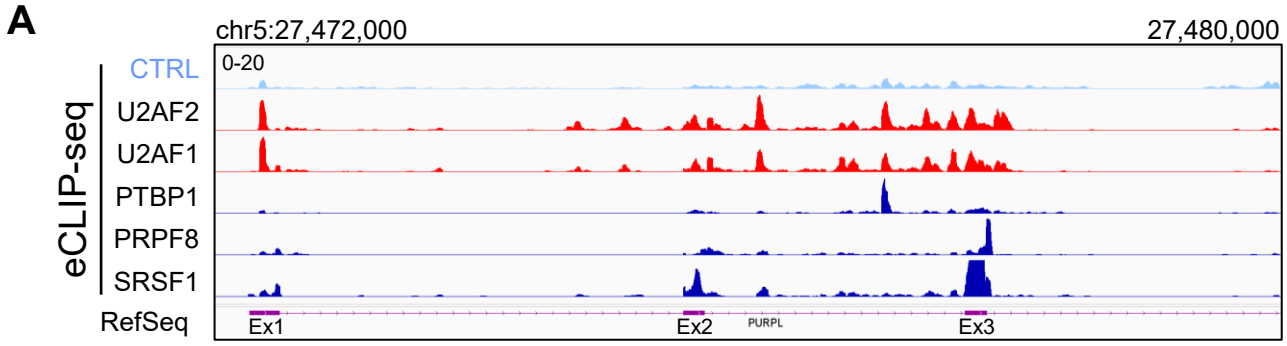


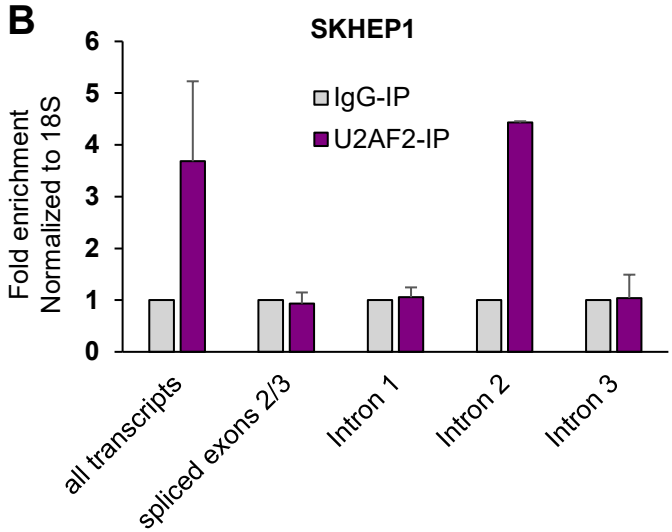


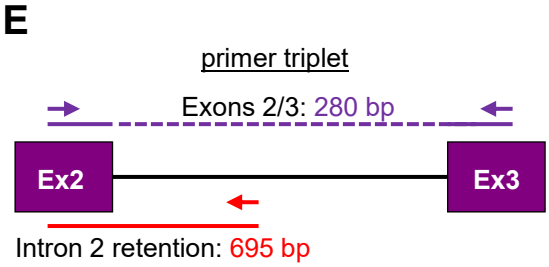
F

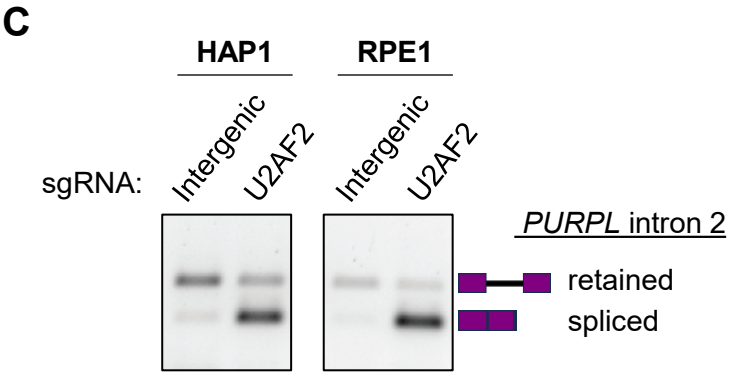
HIT	ΔPIR HAP1	ΔPIR RPE1	Name
U2AF2	-28.42	-23.98	U2 Small Nuclear RNA Auxiliary Factor 2
SNRNP70	-16.05	-3.69	Small Nuclear Ribonucleoprotein U1 Subunit 70
EFTUD2	-14.64	-12.84	Elongation Factor Tu GTP Binding Domain Containing 2
DDX23	-9.49	-6.28	DEAD-Box Helicase 23
SART3	-2.01	-14.01	Spliceosome Associated Factor 3, U4/U6 Recycling Protein

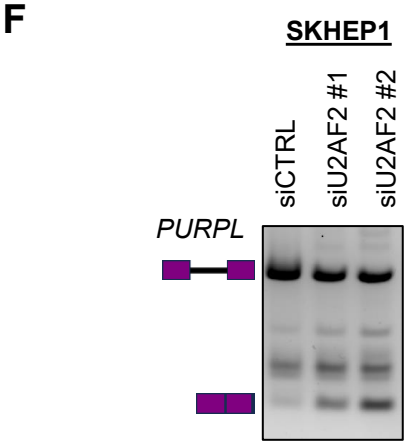
Figure 2

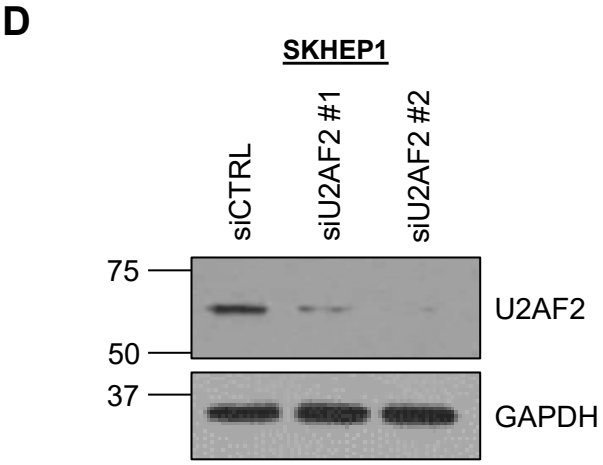

A
chr5:27,472,000
27,480,000
0-20
CTRL
U2AF2
U2AF1
PTBP1
PRPF8
SRSF1
eCLIP-seq
RefSeq
Ex1
Ex2
PURPL
Ex3


B
SKHEP1
IgG-IP
U2AF2-IP
Fold enrichment
Normalized to 18S
0
1
2
3
4
5
6
all transcripts
spliced exons 2/3
Intron 1
Intron 2
Intron 3


E
primer triplet
Exons 2/3: 280 bp
Ex2
Ex3
Intron 2 retention: 695 bp


C
HAP1
RPE1
sgRNA:
Intergenic
U2AF2
Intergenic
U2AF2
PURPL intron 2
retained
spliced


F
SKHEP1
siCTRL
siU2AF2 #1
siU2AF2 #2
PURPL


D
SKHEP1
siCTRL
siU2AF2 #1
siU2AF2 #2
75
50
37
U2AF2
GAPDH

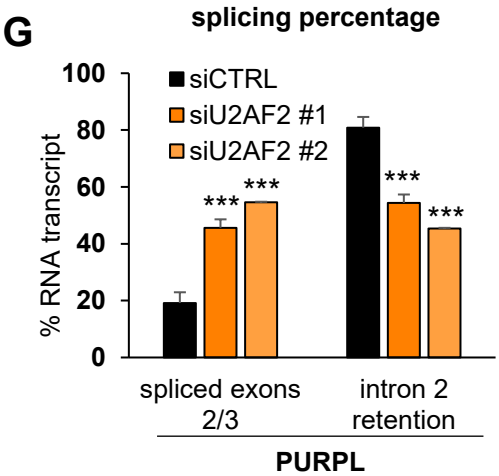
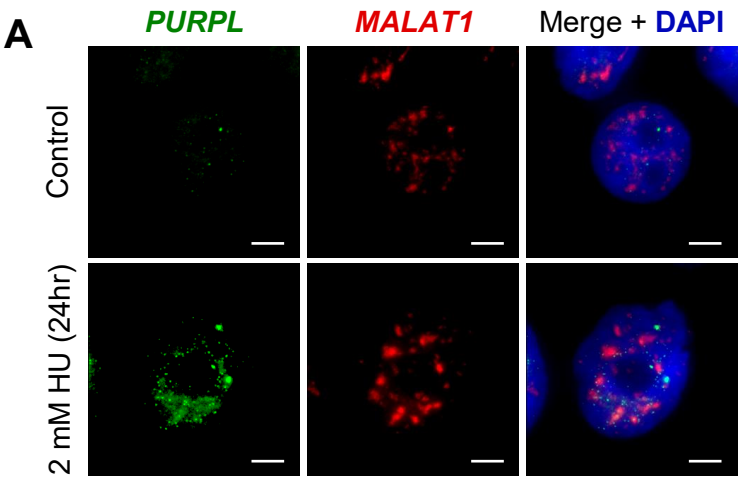
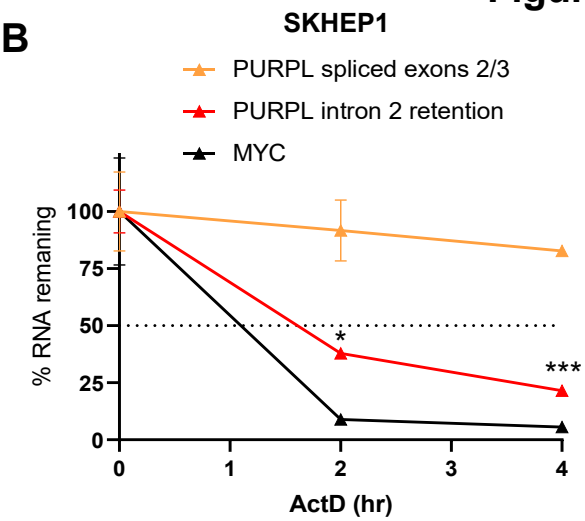
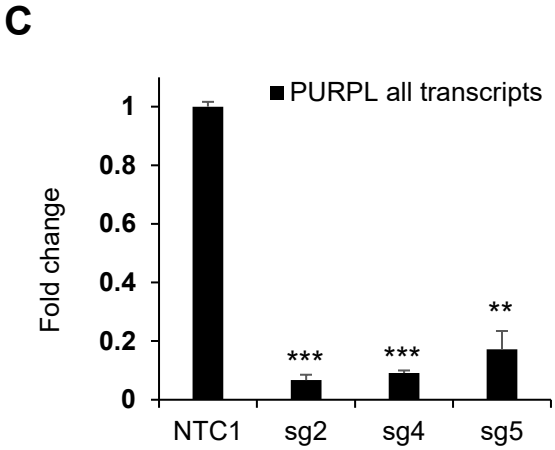
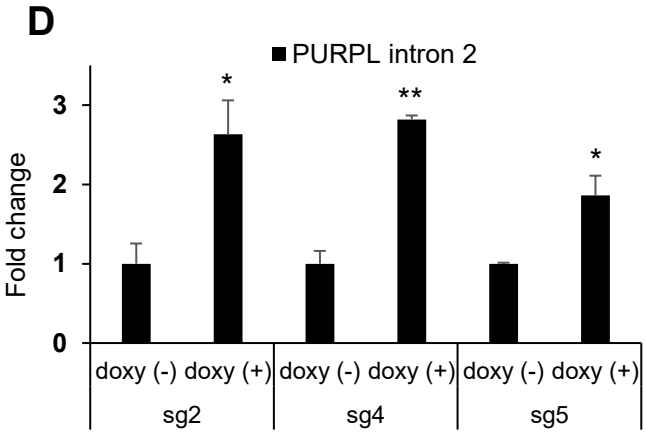

G
splicing percentage
siCTRL
siU2AF2 #1
siU2AF2 #2
% RNA transcript
0
20
40
60
80
100
***
***
***
***
spliced exons 2/3
intron 2 retention
PURPL

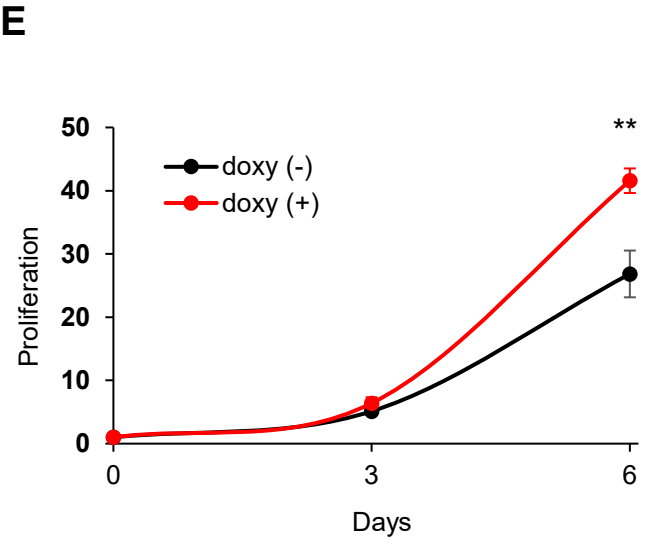
Figure 3


A
PURPL
MALAT1
Merge + DAPI
Control
2 mM HU (24hr)


B
SKHEP1
PURPL spliced exons 2/3
PURPL intron 2 retention
MYC
% RNA remaning
100
75
50
25
0
*
***
0
1
2
3
4
ActD (hr)


C
PURPL all transcripts
Fold change
1
0.8
0.6
0.4
0.2
0
***
***
**
NTC1
sg2
sg4
sg5


D
PURPL intron 2
Fold change
3
2
1
0
*
**
*
doxy (-)
doxy (+)
doxy (-)
doxy (+)
doxy (-)
doxy (+)
sg2
sg4
sg5


E
doxy (-)
doxy (+)
Proliferation
50
40
30
20
10
0
**
0
3
6
Days

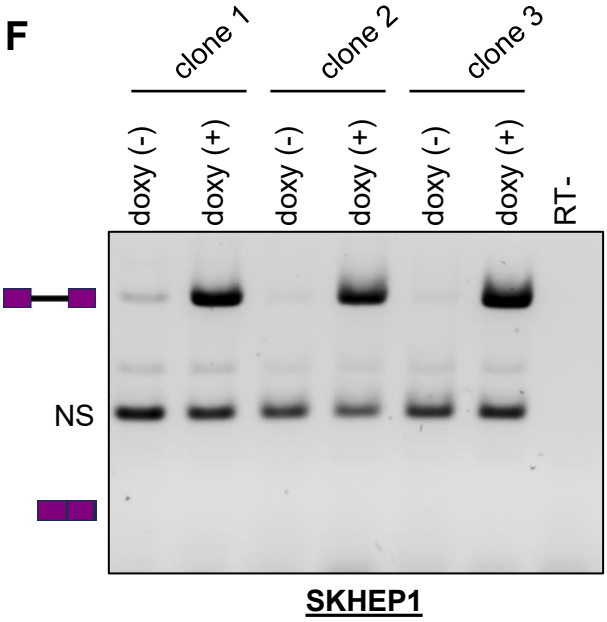
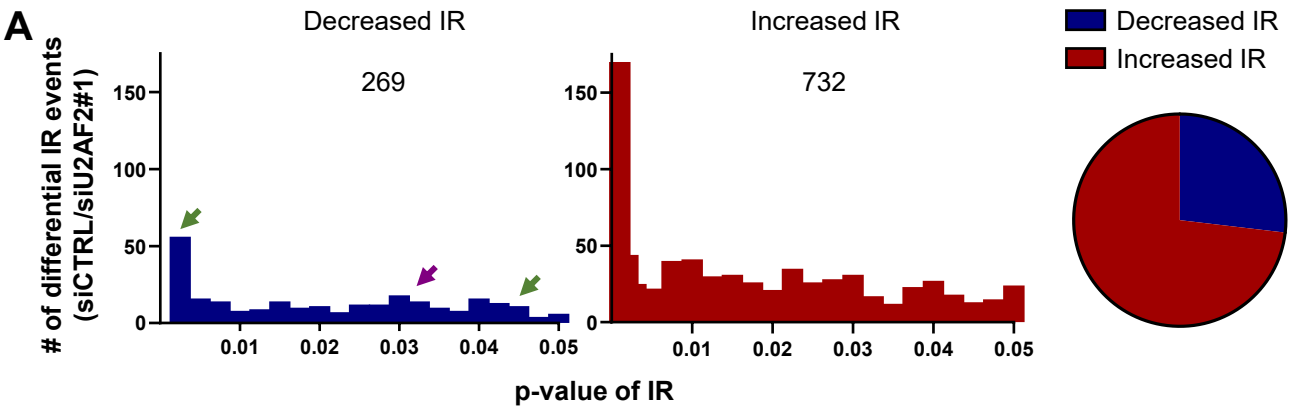
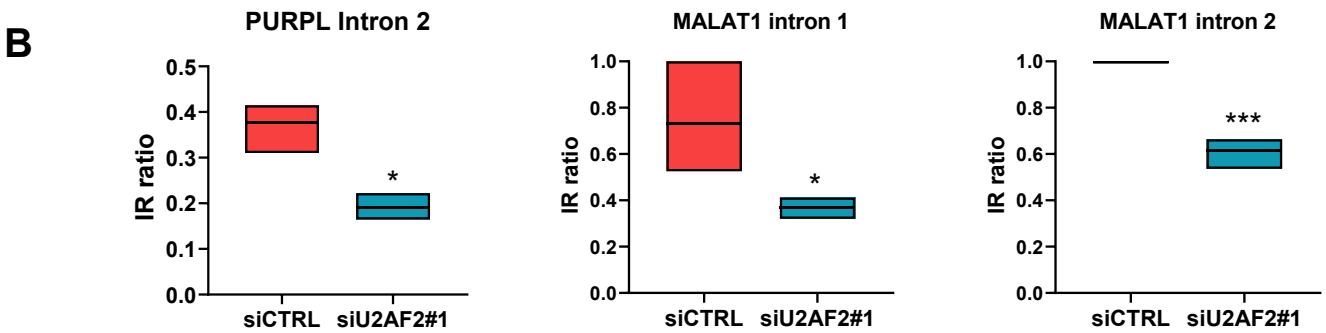
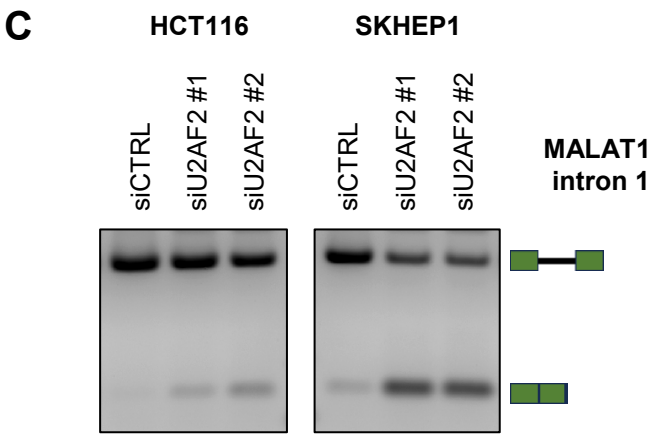
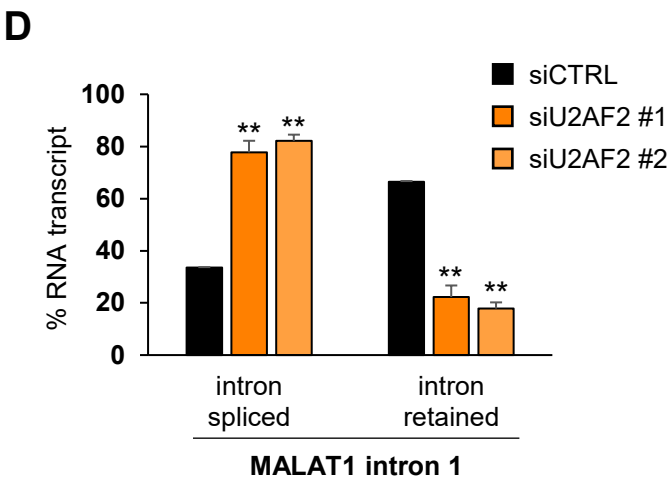

F
clone 1
clone 2
clone 3
doxy (-)
doxy (+)
doxy (-)
doxy (+)
doxy (-)
doxy (+)
RT-
NS
SKHEP1

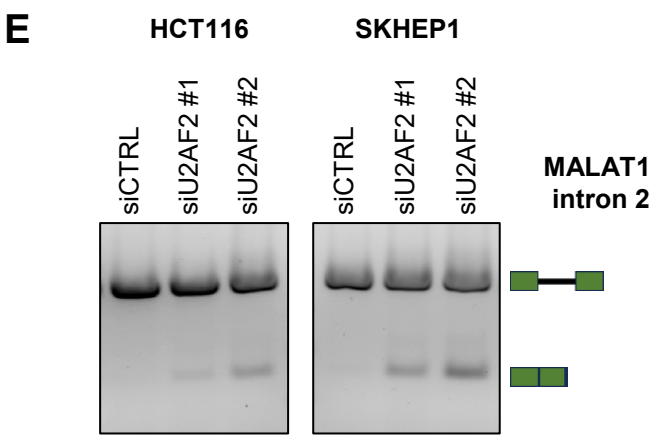
Figure 4


A
Decreased IR
Increased IR
269
732
# of differential IR events (siCTRL/siU2AF2#1)
p-value of IR
Decreased IR
Increased IR


B
PURPL Intron 2
MALAT1 intron 1
MALAT1 intron 2
IR ratio
siCTRL
siU2AF2#1


C
HCT116
SKHEP1
siCTRL
siU2AF2 #1
siU2AF2 #2
MALAT1 intron 1


D
siCTRL
siU2AF2 #1
siU2AF2 #2
% RNA transcript
intron spliced
intron retained
MALAT1 intron 1


E
HCT116
SKHEP1
siCTRL
siU2AF2 #1
siU2AF2 #2
MALAT1 intron 2

F
siCTRL
siU2AF2 #1
siU2AF2 #2
% RNA transcript
intron spliced
intron retained
MALAT1 intron 2

Figure 5

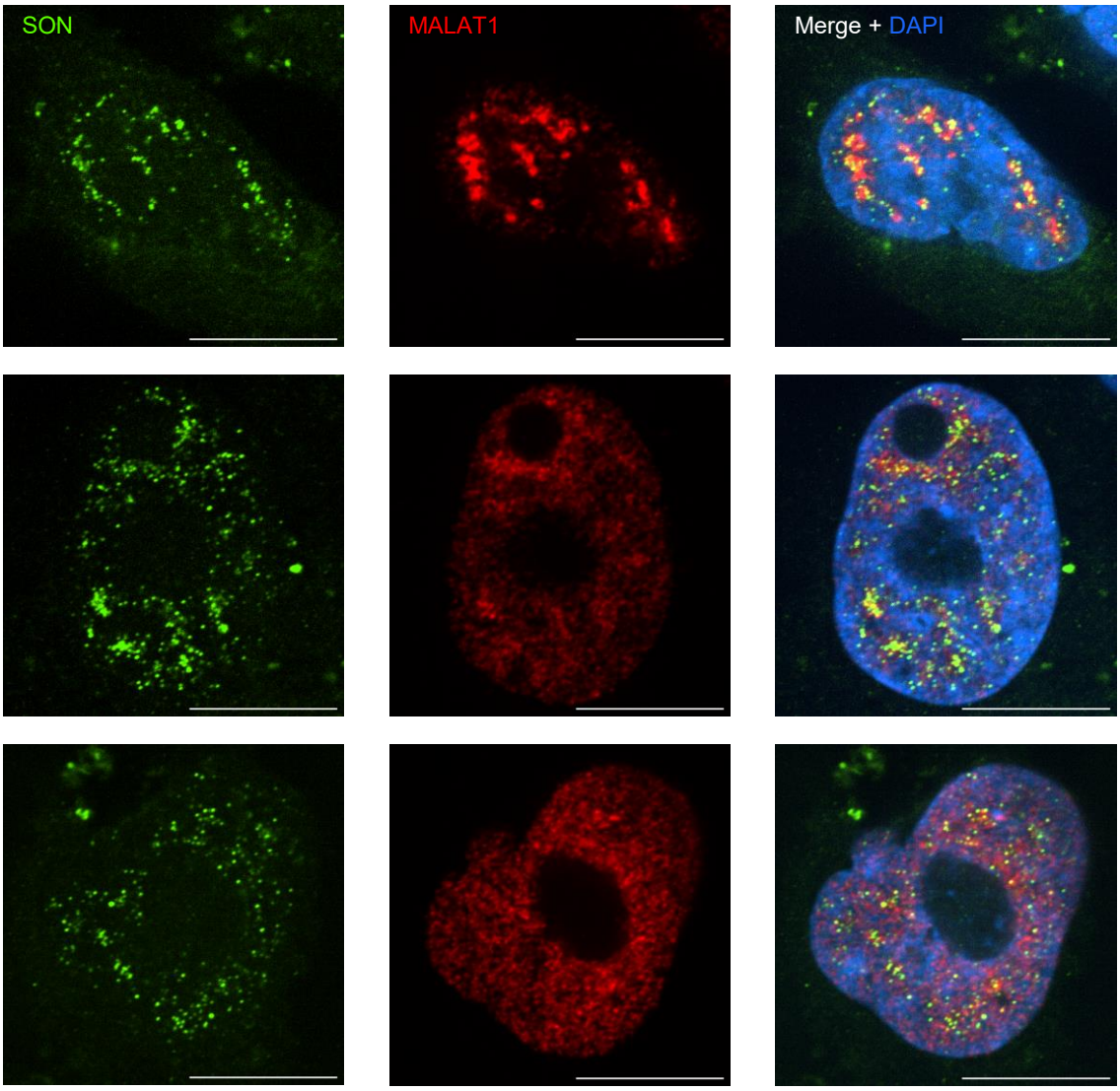
A

SKHEP1

siCTRL

siU2AF2 #1

siU2AF2 #2



B

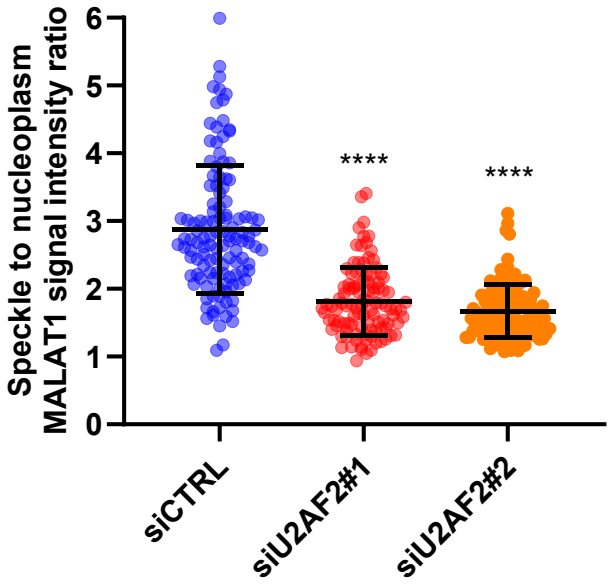
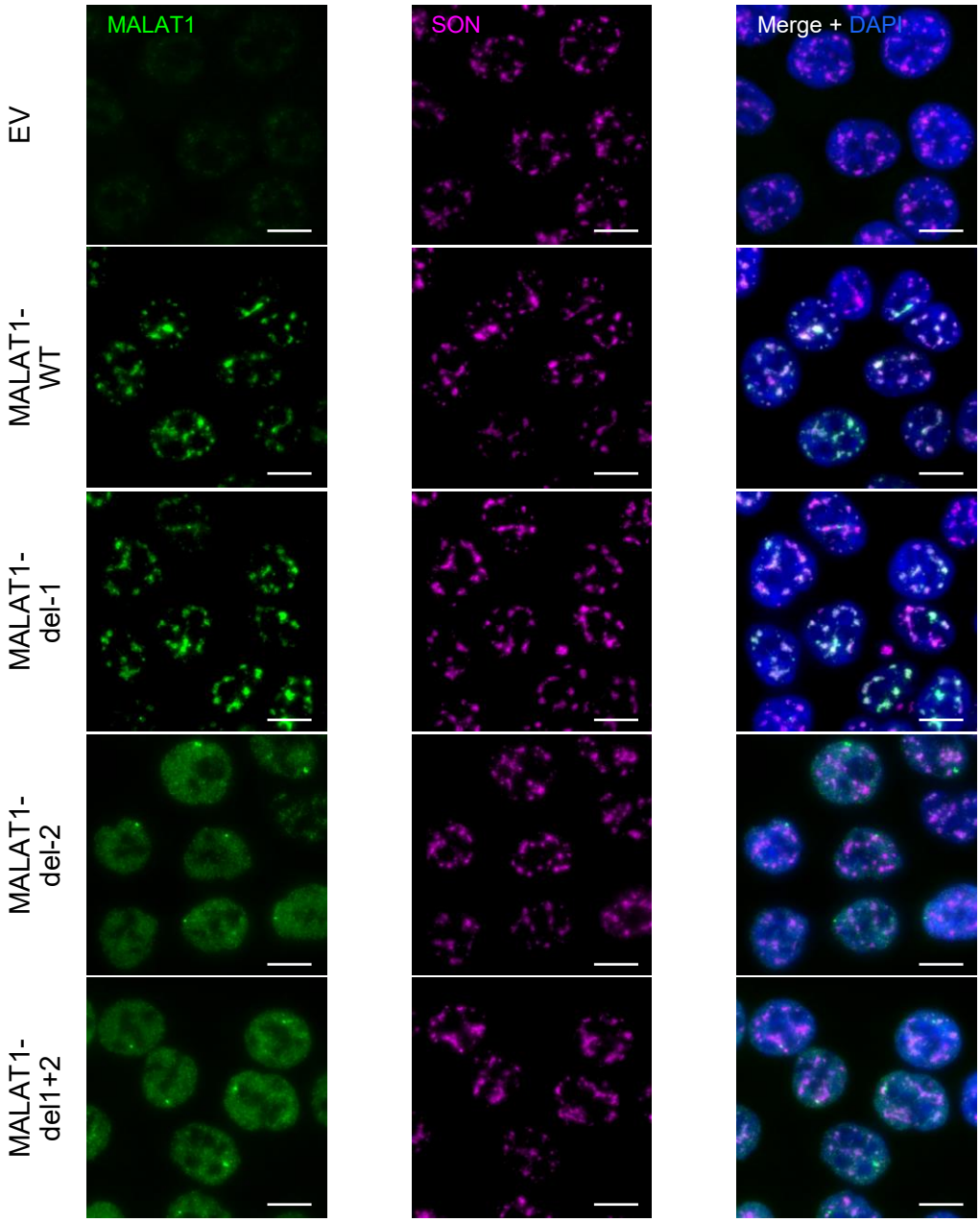


Figure 6

A


MALAT1
SON
Merge + DAPI
EV
MALAT1-WT
MALAT1-del-1
MALAT1-del-2
MALAT1-del1+2

B

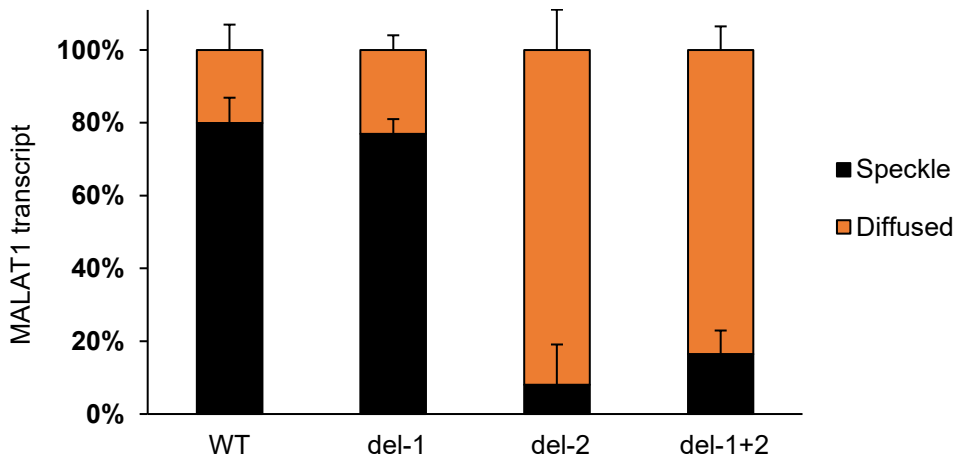
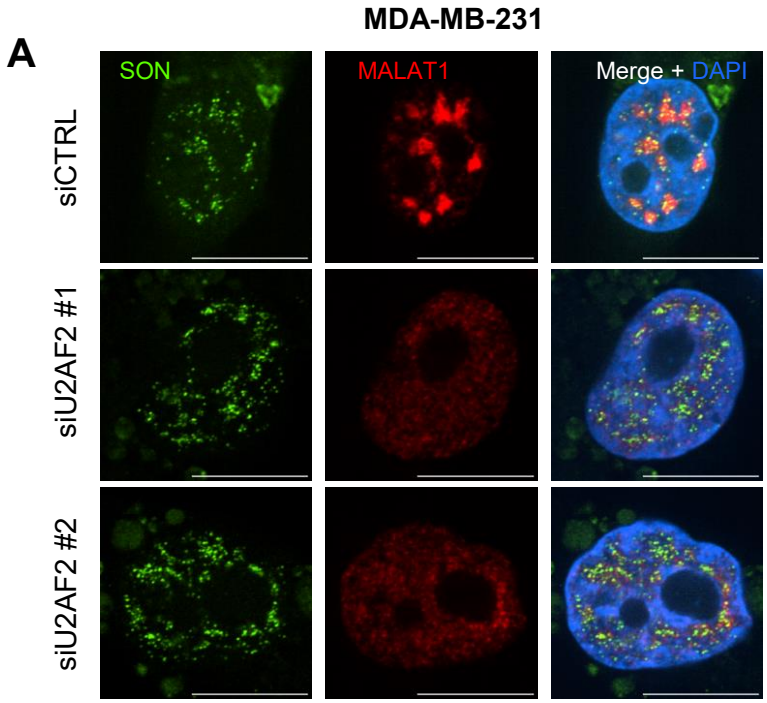
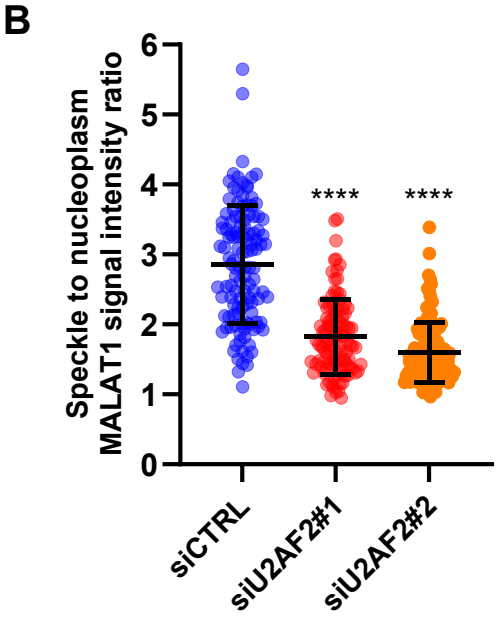
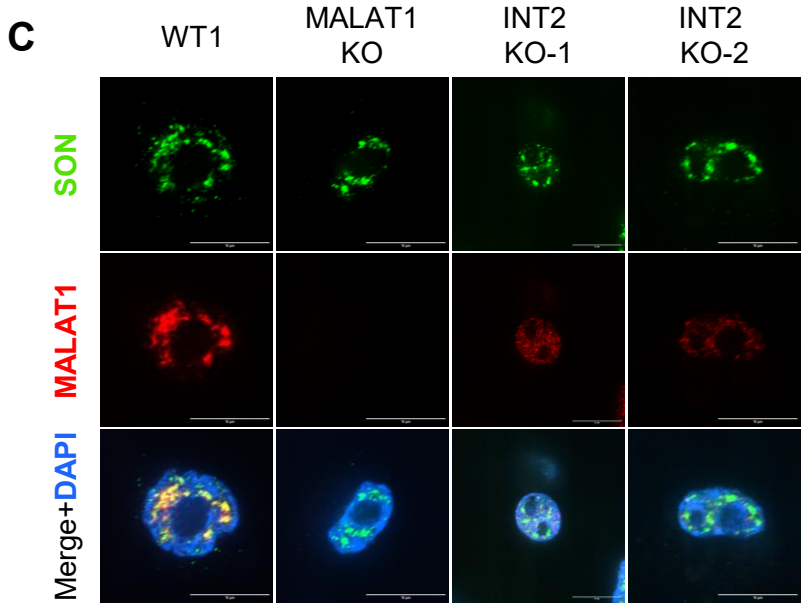
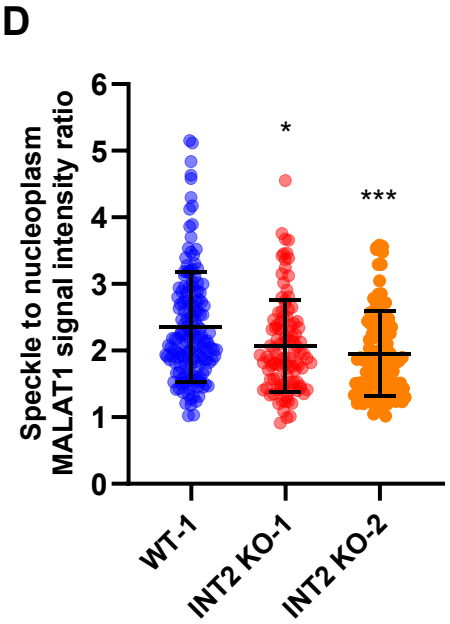

MALAT1 transcript
100%
80%
60%
40%
20%
0%
WT
del-1
del-2
del-1+2
Speckle
Diffused

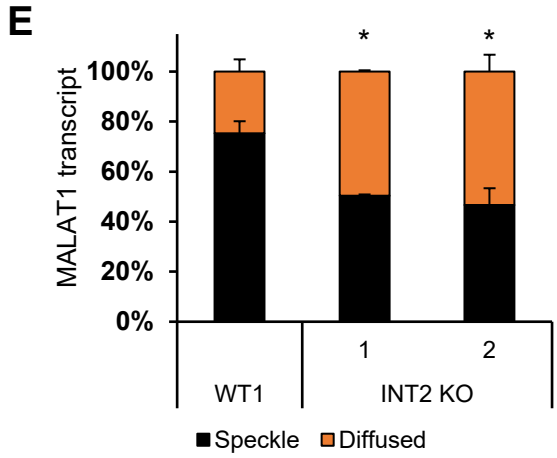
Figure 7

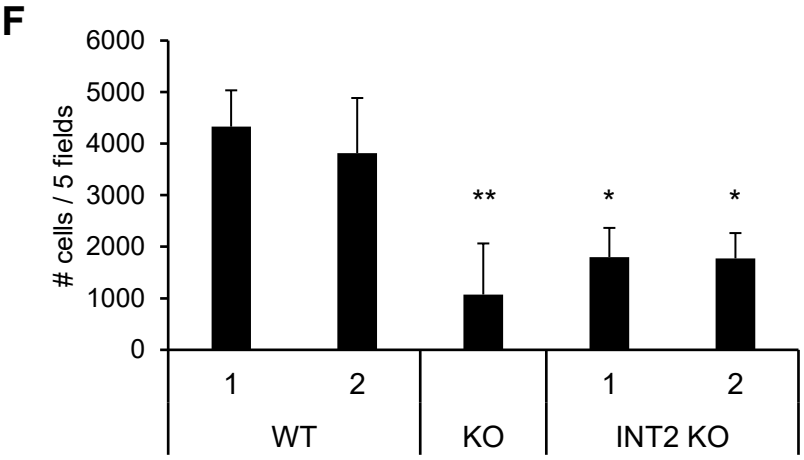

MDA-MB-231
A
SON
MALAT1
Merge + DAPI
siCTRL
siU2AF2 #1
siU2AF2 #2


B
Speckle to nucleoplasm MALAT1 signal intensity ratio
0
1
2
3
4
5
6
****
****
siCTRL
siU2AF2#1
siU2AF2#2


C
WT1
MALAT1 KO
INT2 KO-1
INT2 KO-2
SON
MALAT1
Merge+DAPI


D
Speckle to nucleoplasm MALAT1 signal intensity ratio
0
1
2
3
4
5
6
*
***
WT-1
INT2 KO-1
INT2 KO-2


E
MALAT1 transcript
100%
80%
60%
40%
20%
0%
*
*
1
2
WT1
INT2 KO
Speckle
Diffused


F
cells / 5 fields
6000
5000
4000
3000
2000
1000
0
**
*
*
1
2
1
2
WT
KO
INT2 KO